The Dissertation Committee for Wenting Zou Certifies that this is the approved version of the following dissertation:

Examining Learners' Social Presence in A Massive Open Online Course through Social Network Analysis and Machine Learning

Committee:

Min Liu, Supervisor

Joan E. Hughes

María González-Howard

Cher Ping Lim

Examining Learners' Social Presence in A Massive Open Online Course through Social Network Analysis and Machine Learning

by

Wenting Zou

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2020

Dedication

To Mom and Dad, for all of your love, support and guidance for 30 years. This dissertation would not be possible without you.

Acknowledgements

First and foremost, I owe an immense debt of gratitude to my supervisor, Dr. Min Liu, for inspiring and guiding me through this winding and sometimes rocky five-year journey. Without your patient mentorship and guidance, I would not be typing these words today. Thank you for having such strong faith in me. I would also like to thank Dr. Joan Hughes who provided immensely helpful guidance in navigating the academic world and her efforts in holding the LT family together, which made my graduate school experience at UT much more enjoyable. I am also thankful for the opportunity to learn from my other committee members Dr. María González-Howard and Dr. Cher Ping Lim. I have benefited greatly from your valuable insights and feedback.

I am also deeply indebted to other good friends that have kept me sane and motivated during this journey.

***Alien Rescue* family**: I would like to appreciate all graduate student colleagues in the AR team. Special thanks to Zilong Pan and Chenglu Li, for inspiring and accompanying me in the completion of my dissertation. I also appreciate and cherish the friendship of other current and retired members of the AR family such as Yi Shi, Xin Pan, Jina Kang, Hyeyeon Lee, Sa Liu. Each of you taught me unique things. I greatly enjoyed all the AR meetings, the countless dinner/drinking parties and late night long talks about our struggles. Why are you guys so awesome?

Chirp Lim: I am still immensely grateful for the chance to meet and get to know you in San Antonio four years ago, which opened up so many great opportunities for me to participate in large-scale international studies, and contribute my efforts to bring quality education to the poorest parts of the world. Thank you for believing

in me and helping me expand my research network in Southeast Asia. Your foresight in education and wisdom of life refined my impression of a scholar. You showed me how cool and charismatic a professor can be.

Chris Pan: thank you for being such a great BFF to me. I am lucky to have your strong support in research and in…everything. It gives me great comfort knowing that you'll always have my back when I struggle. I miss the weekends of working, cooking, chatting and drinking with you (even though I suck at drinking). I look forward to nourishing our friendship for many years to come.

Hsiao-Ping Hsu: I am extremely lucky to have you as my best research buddy and closest friend in the PhD journey. Your ambition, determination and perseverance are truly inspiring to me. I miss the countless days and nights brainstorming ideas and hatching plans with you for our project. I am proud of what we have built starting from scratch and the impact our research have made.

Meijia Liu: thank you for reaching out to me when you feel disoriented and anxious in graduate school. And thank you for the warm hugs and encouragement in my vulnerable moments. I am glad that I can always count on you for sensible advice in research and in navigating personal life. Thank you for keeping me sane.

Michael Lee: I consider it a blessing that you came into my life when I needed love and support the most. It's rare to meet someone who had profoundly influenced so many aspects of my life in such a brief journey. You taught me how to truly connect with people with an honest and sincere heart, and not afraid to show my authentic self. You inspired me in so many things that you probably don't know of (even part of this dissertation). I cherish your love, your gentle patience and our sweet memories together. I'm always curious about having a Sheldon-like friend in my life, thank you for making that a reality.

Shanting Chen: your constant encouragement and appreciation of my research always motivate me to aim higher and work harder. Writing, dining and chatting with you are a part of my fondest memories of Austin.

Wenchao Yu: thank you for guiding me onto this academic path. You taught me to work hard, dream big and believe in myself that I am smart and capable enough the change the world. I know this dissertation is far from perfect by your standards, but it's a good start. Thank you for having faith in me. I've grown to be a more mature and resilient person because of you. I look forward to the opportunities to still do something about those world-changing dreams that we once had.

Yi Shi: it is rare to find someone that you get along with on so many levels. Words can't express how much I appreciate having you by my side in this journey. I miss our long walks and talks along the trails by Colorado River in summer nights. Thank you for always looking after me and comforting me with your good foods and kind words.

Yuyang Wang: thank you for your unconditional love for the past five years. I miss having you holding my hands and offering me a shoulder to cry on whenever I struggled. I realized you were the main reason why I never felt lonely in this long journey. I miss all the crazy and exciting adventures we had exploring this new land. You were a horrendous roommate, but a priceless friend. Even though life took us into separate paths, it gives me great comfort knowing that wherever I go, Texas is home. Because you are here.

Lastly, many thanks to this wonder journey filled with joy, laughter, bitterness, frustration, doubt, struggles and love. A PhD is so much more than a degree. It broke me down into my most vulnerable form, and built me back together to become a more resilient, mature, persistent and humble person. As this point, as I look back into the past,

I can still feel the excitement during those quiet midnights sitting in front of my laptop crafting ideas, plans and dreams for the future. Finally, I greatly appreciate the unconditional love and support of my family, for allowing me the freedom to pursue my wildest and adventurous dreams.

Abstract

Examining Learners' Social Presence in A Massive Open Online Course through Social Network Analysis and Machine Learning

Wenting Zou, Ph.D.

The University of Texas at Austin, 2020

Supervisor: Min Liu

Abstract: Low engagement has been a longstanding problem in Massive Open Online Courses (MOOCs). However, engagement is crucial in social learning contexts to increase knowledge construction and achieve meaningful learning outcome. To further understand learners' engagement in MOOC discussion forums, this study focuses on the perspective of social presence, which is defined as learners' ability to project themselves socially and emotionally in a community of inquiry. Social presence is an important factor that has the potential to affect learners' learning experience and outcome. This study took place in the context of a professional development MOOC in the field of journalism. The discussion posts, system log data and survey responses were collected and analyzed. The purpose of this study is to understand the learners' participation patterns in the discussion forums over the six modules of the MOOC, and the relationship between learners' social presence, their positions in the learner network and their learning outcomes.

In terms of data analysis, this study adopted a mixed-method approach to examine the data from both qualitative and quantitative aspects: to qualitatively analyze the posts,

a machine learning supported text classification model was developed and applied to automatically analyze the large-scale text data in the forums; social network analysis (SNA) was used to analyze the characteristics of the learner network and determine learners' centrality (degree, closeness, betweenness and Eigen centrality). Centrality is an important measure because prior studies found it to be an important predictor of learning outcome. Correlation analyses were used to discern the relationship between social presence and learners' centrality, while regression models were built to investigate how learners' social presence and posting behaviors (frequency of posting, average length of posts and day of posting) predict learners' network centrality. Finally, correlation analyses were conducted to understand the association between learners' network centrality and their certificate status, perceived learning and satisfaction. The purpose of using mixed methods is to see in what ways the qualitative nature of the posts and learners' posting behaviors impact learners' positions and influence in the learning community and their learning outcomes.

The findings revealed the evolvement of the learner network in relation to the distribution of social presence throughout the MOOC. The results also showed that social presence indicators such as *Complimenting others, Expressing agreement, Expressing gratitude* and *Disagreement/doubts/criticism* play important roles in learners' centrality in the learner network. Beside social presence, frequency of posting has strong effect in predicting learners' network centrality, while other factors such as the average length of posts and the timing of posting have marginal impact in the prediction. Finally, this study found that learners' network centrality is correlated with their certificate status as well as their overall satisfaction with the MOOC, but not correlated with their perceived learning in the MOOC. This study is among the first efforts in MOOC research to examine the relationship between social presence, learners' network centrality and learning outcomes.

It provides a critical ground for studying content-related interaction and learning community in MOOC forums. The findings inform MOOC learners in terms of how to strategically present themselves in the discussion forums to increase the possibilities of peer interaction and achieve productive learning outcomes. For examples, findings suggest that learners may obtain more central position in the community by posting more compliments, expressing more gratitude, and communicating agreement and disagreement, doubts etc. While for MOOC instructors, this study will potentially inform them how to effectively mediate the discussions and improve learner engagement as a facilitator, such as paying attention to the changes of learner network, identifying central learners, monitoring learners' affective states.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

SIGNIFICANCE OF THE STUDY

With the advent of Massive Open Online Courses (MOOCs) in recent years, there is a tremendous research effort to trying to understand the dynamics of learner behaviors for the purpose of retaining and engaging learners to help them achieve more meaningful learning outcomes (Liu at al., 2019; Poquet & Dawson, 2016; Qu & Chen, 2015; Tseng et al., 2016). Due to the openness and the scale of MOOCs, one of the most distinct features of it is the massive and heterogeneous learners who participate in the courses (Kizilcec, Saltarelli, Reich, & Cohen, 2017). These diverse learners join MOOCs with different motivations, backgrounds, learning strategies and expected goals. For example, a professional development MOOC may have participants from different parts of the world. Sometimes they sign up for a MOOC with a topic that is only remotely related to their own profession, or has no relevance to their jobs at all. For instance, a finance analyst may sign up for a MOOC with a topic in public health. There are consistent participants with an ultimate goal of obtaining the course certificate, while others only participate in specific parts of the MOOC to increase their knowledge in certain areas. As a result, in contrast with traditional courses confined in Learning Management Systems (LMSs), which are often designed for bounded groups with similar backgrounds coming together to take a course for credit on the same commencement and completion dates, MOOCs experienced higher attrition rate due to the heterogeneous nature of the audiences and the asynchrony of participation. During the last five years, a great number of MOOC researchers have been hypothesizing and proving the various factors that link to learners' engagement and performance, such as social interaction and rapport, difficulty of content, motivation, pedagogy, the demographics of participants and so on

(Fidalgo-Blanco, Sein-Echaluce, & García-Peñalvo, 2016; García-Peñalvo, Fidalgo-Blanco, & Sein-Echaluce, 2018; Gütl, Rizzardini, Chang, & Morales, 2014; Stiller & Bachmaier, 2017). Among them, the socio-emotional factor, which captures the management of emotions and the ability to establish positive and rewarding relationships with others (Cohen, Onunaku, Clothier, & Poppe, 2005), gained a lot of attention and researchers found that it has significant impact on learners' success in MOOCs (Lu & Churchill, 2014; Zhang, Yin, Luo, & Yan, 2017). A typical way to examine learners' socio-emotional status is to analyze their participation in discussion forums, where learners express their thoughts, exchange ideas and construct knowledge through social interactions (Al-Rahmi, Alias, Othman, Marin, & Tur, 2018).

According to existing literature in online learning, learners regularly report feeling isolated and alone when taking online courses (Bischoff, 2000; Croft, Dalton, & Grant, 2010; Ludwig-Hardman & Dunlap, 2003). The lack of social connections is amplified in MOOCs where there are hundreds, if not thousands, of learners (Baggaley, 2013). Empirical studies have provided ample evidence that social connections among learners play an important role in their academic performance, resilience, satisfaction, and sense of belonging in their course of study (Poquet & Dawson, 2016). However, it requires high-level skills for educators to design and moderate a learning environment that encourages the cultivation of social rapport among learners which is conducive to learning. As online learning becomes more common in higher education, the design and moderation of social interactions become far more complex for instructors and course facilitators to undertake. And the complexity compounds in the context of MOOCs. The limitation in the functions of discussion forum and instructors' time constraints hinder the effective intervention in students' discussions that could potentially promote meaningful interactions amongst the course participants. From the perspective of learners, the

overwhelming volume of interactions may be perceived as chaotic, hence discouraging learners' effort to navigate and make sense of them, which may later lead to learner frustration, disengagement, and eventually failure (Knox, 2014). To address this problem, there is a need to systematically analyze and understand the social behaviors exhibited by learners. The insights from learners' social patterns allow more timely and effective facilitation of cultivating interpersonal relationships within the learning communities and deepen learners' cognitive engagement.

One way to study learners' social behaviors in online learning settings is through the analysis of social presence. The concept of social presence comes from the Community of Inquiry (CoI) framework developed by Garrison, Anderson, and Archer in the late 1990s (Garrison, Anderson, & Archer, 2001; Garrison & Arbaugh, 2007). Garrison and his colleagues posited that meaningful learning takes place in a CoI through the interaction of three core elements: social presence, teaching presence, and cognitive presence. The first element, social presence, is the ability of participants "to project their personal characteristics into the community, thereby presenting themselves to other participants as 'real people'" (Garrison et. al., 2001, p. 89). The second element, teaching presence, involves instructional management, building understanding, and direct instruction. And the third element, cognitive presence, is "the extent to which the participants in a community of inquiry are able to construct meaning through sustained communication" (Garrison et al., 2001, p. 89). Garrison and his colleagues believed these three elements contribute to a meaningful educational experience. In an earlier stage, researchers examined these three elements individually (Arbaugh & Hwang, 2006; McKlin, Harmon, Evans, & Jone, 2001; Rourke & Anderson, 2002), and social presence received the most attention and interest (Garrison, 2007). This is partly due to a long history, dating back to the 1970s (Lowenthal, 2009), of researchers trying to understand

learning in a social and constructivist context. Even though more recent research has shifted the focus to study the combination of all three elements (Akyol, Vaughan, & Garrison, 2011; Arbaugh, Bangert, & Cleveland-Innes, 2010; Ke, 2010), there is still value in dissecting learners' social presence in isolation of the other two constructs, since social presence itself provides rich contextual information of the social climate in which learning takes place, which warrants thorough examination in relation to learners' engagement and overall performance.

In order to better understand the social processes in which learning occurs, social network analysis (SNA) is becoming increasingly popular in the realm of education research through extracting the patterns of connections among learners. In the context of MOOCs, SNA is often used to untangle the complex learner network and examine the relationship between social network properties and learning outcomes. For example, Poquet and Dawson (2016) applied SNA to understand how social processes unfold in a particular cohort defined by its participants' regularity of forums presence. They analyzed this cohort and its development in comparison to the entire MOOC learner network. Results showed that the cohort, similar to its bounded counterparts in formal online education, could potentially cultivate interpersonal relationships and gradually deepen a shared cognitive engagement. Another study conducted by Yang and her colleagues (Yang et al., 2016) suggested that learners who participated in the forums early on were more likely to complete the course, whereas those who joined later found it hard to form social connections with peers. However, learners' network position does not consistently predict their performance in MOOCs. For instance, Jiang, Fitzhugh, and Warschauer (2014) found a significant correlation between learners' final grades and their network centrality measures (parameters that evaluate how central a node is in a network) in algebra MOOC, but this pattern did not apply to a finance MOOC. Similarly, Joksimović

et al. (2016) found that some centrality measures significantly correlated with the learners' completion and distinction status in two offerings of a programming MOOC, while others were useful in one course but not the other. Another study by Houston, Brady, Narasimham, and Fisher (2017) found that direct learner interactions on the forums are more often correlated with learners' final grades than indirect measures. Taken together, research using SNA to examine learners' experiences in MOOC revealed inconclusive findings regarding how learners' interactions in the network affect their learning experience and outcome. More empirical studies are needed to uncover the social dynamics during the learning processes in online settings, specifically studies focused on what factors contribute to learners' centrality in a learning community, and how that links to their learning outcome.

Besides SNA, content analysis of learners' discussions is another approach to understand the social dynamics during the learning processes. In early years of MOOC research, most of the studies used a quantitative approach to analyze learners' behaviors in terms of their interactions with course components and participation in discussion forums (Brinton & Chiang, 2015; Kloft, Stiehler, Zheng, & Pinkwart, 2014; Shi, Cristea, Toda, & Oliveira, 2020). Recently, research interests have gradually shifted from the quantitative method to a mixed method by taking into account the qualitative nature of students' communication in MOOCs (Atapattu & Falkner, 2016; Rossi & Gnawali, 2014; Wen et al., 2014; Wise, Cui, & Vytasek, 2016). Due to the massive enrollment of MOOCs compared to traditional online courses, there is a need to use automated computational models to tackle the challenge of understanding the large scale of discourse in MOOC forums, which will supplement SNA techniques by adding rich contextual information to the structural patterns of learner interactions (Dowell et al., 2015). While the majority of previous studies focus on how students' network centrality

predicts learning outcomes, this study seeks to probe what kinds of online social presence, as seen through learner engagement in discourse forums, contribute to learners' network centrality, and how that associates with learning outcomes. As a methodological contribution, this study proposes a theoretically grounded computational linguistics model based on the chosen framework of social presence to automate the analysis of students' forums posts. As a theoretical contribution, this study fills the gap of understanding the relationship between social presence, learners' network centrality and learning outcomes. It provides a critical ground for studying content-related interaction and learning community in MOOC forums. The findings will inform MOOC learners in terms of how to strategically present themselves in the discussion forums to increase the possibilities of more peer interactions and achieve productive learning outcomes. While for MOOC instructors, this study will potentially inform them how to effectively mediate the discussions and improve learner engagement as a facilitator.

PURPOSE OF THE STUDY

While online presence is associated with important cognitive processes that precede learning, and affects learner experiences in online environments (Garrison, Anderson, & Archer, 2003), empirical evidence of the impact of social presence on students' learning experience remains inconclusive. In fact, researchers are skeptical about the effectiveness of online courses facilitated by discussion forums in that online environments are unable to provide interaction equivalent to face-to-face discussions (Kohlmeyer, Seese, & Sincich, 2011). Furthermore, given the complexity of interaction emergent from the massive nature of MOOCs, as well as the diverse demographic backgrounds of students, the role of social presence in this particular context remains under-explored and requires further clarification. To this end, this study specifically

investigates one MOOC to gain an in-depth understanding of students' interaction from the perspective of social presence.

Students' social presence is often found in learners' communications in discussion forums, which provides a space for active interactions among learners to deepen their understanding of the course topic and co-construct knowledge by fostering a supportive learning community and stimulating critical discourse. Although the posts generated by learners may enrich our understanding of their social interactions and emotional expressions, which might potentially account for their engagement or dropping out, it is extremely energy and time-consuming to manually analyze all the text inputs in discussion forums when the number of learners reaches a certain magnitude. Therefore, most existing studies rely on a purely quantitative approach to investigate learners' social behaviors in MOOCs, such as using survey to ask learners to report their level of engagement in discussion forums, or counting the frequency of social behaviors including posting, replying, liking posts, subscribing to threads, and following other learners etc. (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014; Liu et al., 2019). Although informative, quantitative index of participation does not directly imply the quality of interaction (Meyer, 2004). In response to this criticism, some researchers have conducted content analysis of thread topics (Gillani, Eynon, Osborne, Hjorth, & Roberts, 2014) or used rule-based algorithms to extract linguistic markers (Wen, Yang, & Rose, 2014) to monitor students' behaviors in discussion forums (e.g., answering questions, self-introduction, complaining about difficulties and exchange of social support, discussing off-task topics etc.). Similarly, while trying to discern learners' emotions from discussion posts in MOOCs, most existing studies established analytic models to automatically detect the types of emotions that are either conducive or detrimental to learning. However, the existing work oftentimes focus on detecting one single dimension of

emotion such as struggling or confusion (Wen, Yang, & Rosé, 2014; Yang, Wen, Howley, Kraut, & Rosé, 2015). Although they somewhat expanded our understanding of learner survival in MOOCs, little effort has been invested in trying to understand how students' social presence affect their engagement in the discussion forums based on an integrative framework. In other words, the coarse-grained classifications used in the existing studies yield limited implications for instructional design in MOOCs to retain learners and help them succeed. This study attempts to thoroughly and systematically explore the social presence of learners in relation to their engagement in the learner network in a MOOC, with the aid of both SNA and the most up-to-date computational algorithms on text classification to automate the analysis of students' forums posts.

RESEARCH QUESTIONS

This study adopts mixed methods to explore both the qualitative (learners' social presence) and the quantitative (learners' posting behaviors) nature of learners' forum posts, in order to understand how these features predict learners' influence in the learner network and affect their certificate status, perceived learning and satisfaction in a MOOC. Specifically, this study aims to answer the following four research questions:

1. What social presence did learners exhibit in the discussion forums of the MOOC?
2. How does the structure of learner network change over the six modules of the MOOC? And how does the learners' social presence differ as the learner network evolves over time?
3. What is the relationship between learners' social presence and their centrality in the learner network? And how do learners' posting behaviors contribute to the prediction of their centrality?

4. How do learners' network centrality correlate with their learning outcomes (measured by certificate status, perceived learning and satisfaction)?

In Chapter 2, prior work on social presence, social network analysis and content analysis that inform this study will be presented. Chapter 3 introduces the research design and methodology, and presents the research questions. Chapter 4 reports the results of data analysis. Finally, Chapter 5 discusses the research findings, the implications for MOOC research and practice, and limitations of this study and offers suggestions for future work.

TERM IDENTIFICATION

Social presence

In this study, social presence is defined as learners' ability to project themselves socially and emotionally in a community of inquiry (Garrison, 2007; Leh, 2001). In computer-mediated online learning contexts, similar to face-to-face context, learners can demonstrate various types of presence during the interactions with peers and the instructor(s). For example, asking questions, offering suggestions to others, expressing emotions, sharing external resources, referencing the message of peers etc. This study examines the link between different types of social presence and learning, specifically as measured by certificate status, perceived learning and satisfaction.

Learner network

Learner network refers to the totality of learners' connections with all other peers and the instructors within the learning community as exhibited through their communications in the discussion forums in a MOOC (Wasserman & Faust, 1994).

Network centrality

Network centrality describes how central a node is in a network, which indicates the social influence of that node (Wasserman & Faust, 1994). In this study, each learner is regarded as a node in the network. Learners' network position is measured by four centrality parameters that are commonly used in SNA, namely, degree centrality, closeness centrality, betweenness centrality and Eigen centrality. The definitions and calculation methods of these centrality parameters are provided in Chapter 2.

Network density

The density of a network is defined as the number of ties among all participants in a network divided by the maximum number of all possible ties (Wasserman & Faust, 1994). Thus, the density of a network is at the maximum when all the participants are connected to each other. The density value of a network varies between 0 and 1. In a dense network, participants are more likely to mutually influence each other, and information that circulates in the network is distributed among many participants (Wasserman & Faust, 1994). Thus, a dense network can mediate peer interactions in a meaningful way. It can offer participants more possibilities to pursue pedagogically valuable discourse; it may also facilitate a collective responsibility to enhance knowledge construction and distribution of expertise among participants. This study examines how learners' social presence changed in relation to the evolvement of the learner network, which can be partly described by network density.

Modularity

Modularity is one measure of the structure of networks. It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). Groups are divided based on the density of connections between the nodes. Nodes within the same group generally have denser connections than with nodes

in other groups (Newman, 2006). Modularity describes the different structures of a network. This study examines how learners' social presence changed over the six modules as the structure of the learner network changed over time.

Natural Language Processing

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence that concerns the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data ("Natural language processing," 2020, para. 1). This study uses NLP techniques to build a text classifier to automate the analysis of learners' posts in the discussion forums.

Chapter 2: Review of Literature

This chapter first discusses prior studies about social presence, highlights the importance of it in online learning settings, then reviews the studies of using SNA to understand learners' online interactions, and finally introduces content analysis of learners' online discussions. The implications and limitations of previous research findings are also presented to guide the design of this study.

SOCIAL PRESENCE

The use of online delivery as an education and training tool continues to expand across a variety of settings. Previous research demonstrates that students in online settings do not participate at the same level of consistency as students do in traditional face-to-face settings (Rovai, 2002). While there are still multiple factors to explore that may influence students' learning pace and patterns in online environments, it is reasonable to assume that learners would want to find ways to make the interaction online as enjoyable and effective as possible. One way to enhance the learning experience online is to engage in social interactions with peers and instructors. Compared to face-to-face learning, establishing social relations with others online is no doubt more challenging and requires a persistent and deliberate effort on the part of the learners, course facilitators, and instructors.

To provide online learners with a sense of presence similar to that in face-to-face instruction, it is crucial to offer interpersonal communication opportunities for students to socially engage with the teacher and peers. In fact, the importance of online presence has been well-documented by earlier studies (Garrison & Cleveland-Innes, 2005; Richardson & Swan, 2003; Swan & Shih, 2005), and pedagogical practices capitalizing on interactive communication technologies are well documented in the literature (Cunningham, 2015).

For example, many online courses integrate social media into their delivery, while others incorporate a wide range of asynchronous facilities such as online discussion forums, wiki, and blog systems (Dabbagh & Kitsantas, 2012; Ke, 2010). In the context of MOOC, the most common way to encourage social interaction among learners is by using the discussion forums. Students' social presence, as reflected in their conversations in the discussion forums, and associated with important cognitive processes that precede learning, is therefore attracting increasing attention in the research community to further understand how the dynamics of social interaction affect learning experience and performance.

The history and definitions of social presence

The roots of social presence can be traced back to the concept of immediacy, which is grounded in the Implicit Communication Theory proposed by Albert Mehrabian (1969). He defined immediacy as communication behaviors that "enhance closeness to and nonverbal interaction with another" (Mehrabian, 1969, p. 203). Immediacy has been linked to the motivational trait of approach avoidance in that, "people approach what they like and avoid what they don't like" (Mehrabian, 1981, p. 22). The concept of social presence was built upon the concept of immediacy, and many applications of social presence today are found within the literature in the field of communications. For example, in studying face-to-face, audio, and closed-circuit television encounters, Short, Williams, and Christie (1976) defined social presence as the "degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationships" (p. 65). To distinguish verbal interaction from nonverbal interaction, Gunawardena and Zittle (1997) proposed another concept, intimacy, in additional to immediacy. They suggested that intimacy and immediacy are two different concepts

associated with social presence in which intimacy is dependent on nonverbal factors, including physical distance, eye contact, and smiling. Immediacy is a "measure of the psychological distance that a communicator puts between himself or herself and the object of his/her communication" (Gunawardena & Zittle, 1997, p. 9).

In contrast, from an instructional communication perspective, Kearney, Plax, and Wendt-Wasco (1985), Gorham (1988), and Christophel (1990) provided some of the early discussions of the concept of social presence, defining it as "teacher immediacy" in the classroom. Behaviors that create immediacy include both verbal and nonverbal actions such as gesturing, smiling, using humor and vocal variety, personalizing examples, addressing students by name, questioning, praising, initiating discussion, encouraging feedback, and avoiding tense body positions (Hackman & Walker, 1990). Rourke, Anderson, Garrison, and Archer (1999) placed more responsibility on the learners when they described social presence as the ability of the learners to socially and affectively project themselves in communities of inquiry.

Additionally, other researchers have offered the following interpretations of the social presence: "the feeling that others are involved in the communication process" (Whiteman, 2002, p. 6); "the degree to which a person feels 'socially present'" (Leh, 2001, p. 110); "the degree of person-to-person awareness" (Tu, 2000, p. 1662); "the sense of being present in a social encounter with another person" (McLellan, 1999, p. 40), and "the degree to which participants are able to project themselves affectively within the medium" (Garrison, 1997, p. 6). However, Gunawardena and Zittle (1997) put it most simply by stating that social presence is "the degree to which a person is perceived as a 'real person' in mediated communication" (p. 9). Drawn from the commonalities of all the above-mentioned descriptions, this study defined social presence as "learners' ability to project themselves socially and emotionally in a community of inquiry". According to

Rourke et al. (1999), The function of social presence is to support the cognitive and affective objectives of learning. Social presence supports cognitive objectives through its ability to instigate, sustain, and support critical thinking in a community of learners. It supports affective objectives by making the group interactions appealing, engaging, and thus intrinsically rewarding, leading to an increase in academic, social, and institutional integration and resulting in increased persistence and course completion (Rourke et al., 1999).

Social presence in the Community of Inquiry (CoI) framework

While social communication and interaction are essential for students to feel connected and to form interpersonal relationships, interaction alone does not guarantee student engagement in the process of cognitive inquiry, nor does it guarantee that cognitive presence is automatically in place (Garrison & Cleveland-Innes, 2005). Socio-cognitive approaches to online learning posit that online presence is a complex construct comprising a multitude of elements in different dimensions, including teaching presence and cognitive presence, in addition to social presence. Furthermore, these elements do not function independently, but rather, there is an interplay among them which forms many intersectional categories that function concurrently to form an integral whole to achieve the full potential of online learning outcomes (Akyol & Garrison, 2008; Garrison & Arbaugh, 2007; Garrison & Cleveland-Innes, 2005). The relationship of these intersectional categories is described by Garrison (2007) in a framework known as the model of Community of Inquiry (CoI) (see Figure 1). A CoI integrates social, cognitive, and teaching presence at the core of online learning experience. The CoI framework was initially constructed on the premise that effective online learning requires the development of community, in which higher order learning occurs when the students

combine their personal experience with shared worlds of experience through interaction with the instructor and peers. The framework aims at establishing an online environment that goes beyond a social community for general social exchange and low-level cognitive interactions, and emphasizes the cultivation of higher-level learning (Akyol & Garrison, 2008; Garrison & Akyol, 2013). At the operational level, a CoI integrates the instructor's role in course design and facilitation, the learners' sense of community and belonging, and their cognitive engagement with the course content (Garrison et al., 2003). It could therefore be used as a theoretical guide to assess different educational approaches and strategies in facilitating a community of inquiry (Akyol et al., 2011).

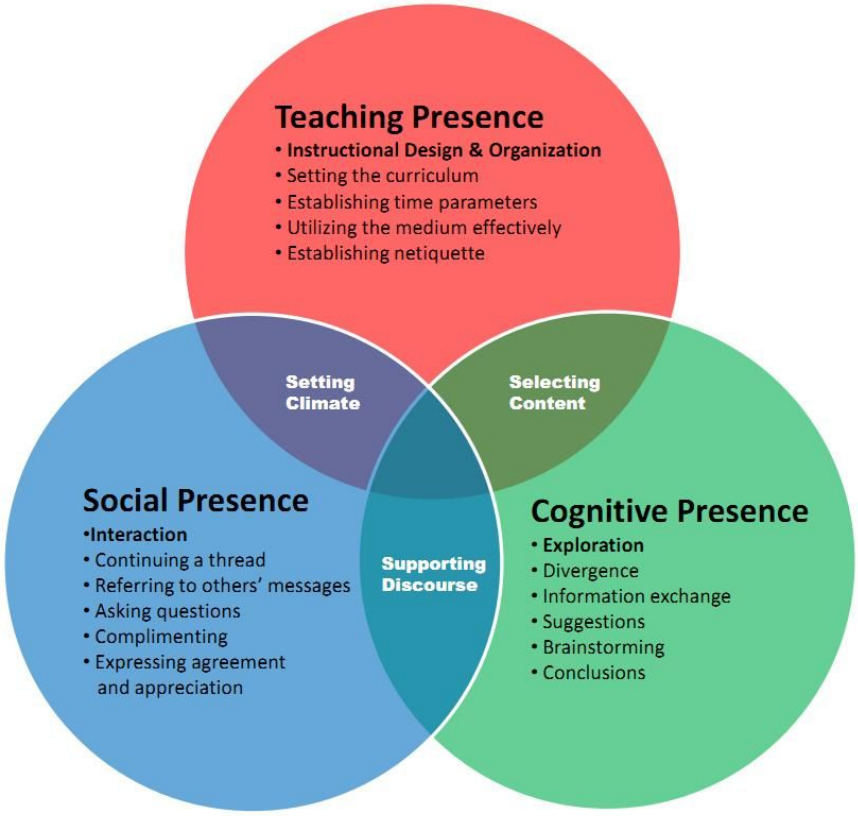


Figure 1: The Model of Community of Inquiry

Although social, cognitive, and teaching presence are equally important components in the CoI framework, this study focuses only on the analysis of social presence in relation to students' learning experience in MOOC. The reasons are: 1) social presence itself provides rich contextual information of the social climate in which learning takes place, which warrants thorough examination in relation to learners' engagement and overall performance (Leh, 2001; Tu, 2000; Whiteman, 2002); 2) the amount of input from instructors was minimum in this MOOC, which renders the analysis of teaching presence limited and unreliable; 3) analysis on the cognitive aspect of students' posts relies heavily on the course topic, which is complicated and challenging to conduct automatic analysis by a text classifier.

Operationalization of social presence in CoI

Social presence is arguably one of the most studied constructs in online learning. Within the CoI framework, social presence reflects the ability of learners to project themselves socially and emotionally, thereby representing themselves as "real" people in a friendly and supportive learning environment (Garrison & Arbaugh, 2007). Social presence can be seen as functional in making cognitive presence more effective, as it provides learners with a relationship-fostering environment for meaning negotiation, collaborative knowledge construction, and critical thinking (Caspi & Blau, 2008; Garrison, Cleveland-Innes, & Fung, 2010; Kozan & Richardson, 2014; Szeto, 2015). As a measure of the feeling of the community with which learners identify themselves in online environments, social presence can be analyzed from three aspects, including *affective expression, open communication,* and *group cohesion* (Akyol & Garrison, 2008) (see Table 1). Social cognitive theory (Bandura, 1989) posits that individuals initiate and regulate their learning to achieve desirable learning outcomes. Interaction with peers and

the situated environment contributes to the development of one's cognition, affect, and behavior. While the dynamic nature of social presence is difficult to capture, studies have highlighted the importance of *affective expression* in establishing a climate for learning before *open communication* and *group cohesion* could develop (Akyol & Garrison, 2008).When provided with a trusting environment, learners can develop interpersonal relationships with other members of the community (Garrison et al., 2010), in which knowledge is socially constructed rather than transmitted or discovered, because increased opportunities for peer learning and interaction allow for the development of rich and elaborate thinking and knowing, which in turn contributes to students' learning at a deeper level (Bransford, Brown, & Cocking, 2000). The role of social presence in establishing a trusting climate was confirmed by Caspi and Blau's (2008), which examined the three categories of social presence and their relations to perceived learning. Their analysis indicated that social presence affords learning by creating a convenient climate.

Table 1: Indicators of Social Presence Devised by Rourke et al. (1999) and Revised by Shea et al. (2010)

Themes	Social presence indicators
Affective Expression	Expressing emotions, Self-disclosure, Expressing values, Unconventional emotion expression, Use of humor
Open Communication	Continuing a thread, Quoting from others' messages, Referring explicitly to others' messages, Asking questions, Complimenting, Expressing appreciation, Expressing agreement, Expressing disagreement, Personal advice
Group Cohesion	Vocatives, Addresses or refers to the group using inclusive pronouns, Phatics, salutations and greetings, Social sharing, Course reflection

Social Presence in online learning

Since the introduction of MOOCs, it has been hailed as a game changer for democratizing higher education by creating free, global, online access to courses from elite universities (Dillahunt, Wang, & Teasley, 2014). However, emerging evidence suggests that most global providers of MOOCs failed to meet this goal because they primarily reached educated learners from affluent countries, potentially widening educational disparities (Kizilcec et al., 2017). For example, among global providers like Coursera, edX or FutureLearn, there are gaps in participation and persistence among learners with relatively low levels of educational attainment and from less-developed countries (Chuang & Ho, 2016; Rohs & Ganz, 2015). For learners with low levels of English proficiency, linguistic and cultural factors further compound the challenges of online learning (Liu, Liu, Lee, & Magjuka, 2010; McLoughlin & Oliver, 2000). With large numbers of highly diversified learners from different geographic locations and personal backgrounds, which are rarely seen in traditional learning environments, researchers identified more demographic variables that affect student success in MOOCs, such as learners' age, prior online learning experience, prior knowledge, educational attainment, and professional status (DeBoer, Stump, Seaton, & Breslow, 2013; Kennedy, Coffrin, De Barba, & Corrin, 2015; Morris, Hotchkiss, & Swinnerton, 2015). In such a unique learning setting, it is crucial to understand the online presence and social dynamics of linguistically and culturally diverse learners in order to better support those who are struggling and have higher risk of dropping out.

Social learning through the use of new technologies has increasingly been adopted as a pedagogical strategy in higher education to harness the emancipatory power of space and interactions outside the physical classrooms (Ryan & Tilbury, 2013). The overall goal for creating social presence in any learning environment, whether it be online

or face-to-face, is to create a level of comfort in which people feel at ease around the instructor and the other participants. Without this goal being achieved, the learning environment can turn to one that is not fulfilling or successful for the instructors and the learners. As Whiteman (2002) states, "People feel more comfortable around us when they believe we share a kinship and common values" (p. 8). When the environment is lacking social presence, the participants see it as impersonal and, in turn, the amount of information shared with others decreases (Leh, 2001).

Since the concept of social presence was first linked to online learning, researchers and practitioners have been reconceiving not only what social presence is, but also the particular role it plays in online learning (Annand, 2011; Gunawardena, 1995; Lowenthal, 2010). Social presence has been shown to influence a variety of factors in students' learning experiences. More typically, social presence can influence students' satisfaction (Cobb, 2011; Gunawardena et al., 2001; Gunawardena & Zittle, 1997; Hostetter & Busch, 2006; Kang, Liew, Kim, & Park, 2014; Strong, Irby, Wynn, & McClure, 2012), perceived learning (Arbaugh, 2008; Cobb, 2011; Kang & Im, 2013; Richardson & Swan, 2003; Swan & Shih, 2005), and actual academic performance (Hostetter & Busch, 2013; Joksimovic, Gasevic, Kovanovic, Riecke, & Hatala, 2015; Picciano, 2002). Ultimately, social presence research underscores the importance of encouraging social interaction as a means to engage learners in critical thinking and higher-level learning (Garrison & Akyol, 2013).

Social presence and student satisfaction

As reported by Shin (2002), much of the research to date has looked at the relationship between the varying extent of social presence and the level of student satisfaction. Gunawardena and her colleagues have produced probably the most extensive

body of empirical research related to social presence and its influence in online environments. In one study, Gunawardena and Zittle (1997) examined the influence of social presence as a predictor of satisfaction within computer-mediated conferencing (CMC) environments. Defining satisfaction as the value of the CMC in facilitating learning for the students, through regression analysis, they found that social presence accounted for 58% of variance in student satisfaction. In a later study, Gunawardena, Nolla, and others (2001) found that social presence facilitates the building of trust and self-disclosure within an online learning context. Likewise, Hostetter and Busch (2006) found that similar levels of social presence could be generated between f-2-f and online course settings. In particular, through regression analysis, they found that 40% of the variance in learner satisfaction could be explained by social presence. Similarly, Strong, Irby, Wynn, and McClure (2012) assessed students' perceptions of the learning environment, social presence, and satisfaction in online agricultural education courses. They found that social presence and the learning environment accounted for 26% of the variance in student satisfaction. These conclusions align with the findings from other researchers, such as Cobb (2011) and Kang, Liew, Kim, and Park (2014), that social presence plays a crucial role in learners' satisfaction of online learning.

Social presence and perceived learning

Richardson and Swan (2003) demonstrated with their correlational study that students who perceived a high level of social presence in an online course were not only more satisfied with their instructor, but also perceived they learned more than students who reported low social presence. Swan and Shih (2005) conducted a mixed-methods study and found significant correlations between perceptions of social presence (peers and instructors) and perceived learning, as well as between the perceived presence of

instructors' and satisfaction with instructors. Cobb's (2011) work on nursing education found that social presence was highly correlated to both student satisfaction and perceived learning. Using multivariate regression, Cobb found that social presence accounted for 36% of the variance in perceived learning. Arbaugh (2008) examined 55 online MBA courses to determine if social presence, cognitive presence and teaching presence could predict students' learning outcomes. He found that social presence was positively associated with students' perceived learning. Similarly, Kang and Im (2013) conducted multiple regression analyses to determine the factors in learner-instructor interaction that predicted learners' perceived learning and satisfaction in online courses. They found that factors related to instructional interaction significantly predicted learners' perceived learning achievement. In summary, most prior studies point to the conclusion that the level of social presence has positive associations with learners' perceived learning in online courses.

Social presence and students' academic performance

Only a few studies have examined social presence in relation to traditional academic performance, or grades. Picciano's (2002) early study examined the impact of interaction and social presence on performance outcomes. After breaking students into three social presence groupings (low, moderate, and high), Picciano compared mean scores for both a written assignment and an examination, and found that students' perceptions of social presence were not a statistically significant predictor for performance on the examination. However, it was a significant predictor for performance on the written assignment. Similarly, Hostetter and Busch (2013) adopted a content analysis approach on the graded discussion postings, a social presence survey, and the Classroom Assessment Technique (CAT) which involved a written assignment as a

measure of academic performance. A regression analysis revealed that students with higher levels of social presence also performed better on the CAT. More recently, using an experimental design approach, Joksimovic et al. (2015) examined the graded online discussion postings. With the treatment groups reporting higher social presence level, the researchers found that certain social presence indicators (i.e., continuing a thread, complimenting, and expressing appreciation) were significant predictors of student academic performance, in this case course grades. This led them to conclude that "the ability of a student to project himself within an online learning community is also a significant predictor of academic performance" (p. 13). They also concluded that instructional design and the inclusion of support for meaningful interactions, which enabled deeper social presence interactions, are critical to improve students' academic performance outcomes.

Social presence in MOOC context

Although the importance of social presence in online learning settings has been well documented (Joksimović, Gašević, Kovanović, Riecke, & Hatala, 2015; Picciano, 2002; Rovai, 2002), the majority of research in the area of social presence is situated within the formal education context. Only a few studies examined learner perceptions of social presence in MOOCs. Kilgore and Lowenthal (2015), for example, found that MOOC participants "were able to experience social presence first hand and that social presence can be established in large online courses" (p. 398). In contrast, Damm (2016) demonstrated that most of their MOOC learners either disagreed that social presence was established in MOOCs, or marked social presence as a non-applicable aspect for their course evaluation. In the context of MOOCs, learners may have very different engagement patterns compared to formal online courses with synchronous participation

by a bound cohort of students. In fact, empirical research of MOOC forums has offered substantial evidence that there are distinct participation patterns in MOOC forums, for example, a small group of learners participate persistently in the forums, whereas a large number of learners participate intermittently, meaning that they engage and disengage randomly (Coffrin, Corrin, de Barba, & Kennedy, 2014; Ferguson & Clow, 2015; Hecking, Chounta, & Hoppe, 2016; Kizilcec, Piech, & Schneider, 2013). Moreover, viewing without posting has also been found to be the activity most characteristic of MOOC forums users (Bergner, Kerr, & Pritchard, 2015). Despite the clear differences between the MOOC context and online courses of smaller scale within a bounded cohort, as well as a high possibility that social presence may unfold differently in MOOCs, research exploring social presence in MOOCs has not addressed these differences methodologically or conceptually.

Summary

In summary, raising social presence in online environments may help create quality learning experience that are conducive to students' satisfaction, perceived learning and actual performance. Social presence in learning leads to inclusion (the need to establish identity with others), control (the need to exercise leadership and prove one's abilities, and affection (the need to develop connection with people) (Whiteman, 2002). High levels of social presence create a learning environment that is perceived as warm, collegial, and approachable for all involved (Rourke et al., 1999). An additional benefit of social presence, according to Rourke et al. (1999), is its ability to instigate, sustain, and support cognitive and affective learning objectives by making group interactions appealing, engaging, and intrinsically rewarding. Despites these purported benefits, the existing studies that examined social presence are limited in bounded learning contexts

instead of in MOOCs. Also, the current analyses of social presence are not in relation to students' engagement in the learner network in MOOCs. Intuitively, the types of social presence learners exhibit in a community may have more direct impact on learners' connection and relationship with their peers, which can be reflected in their status/position in the learner network. However, there is a scant of literature addressing the association between learners' social presence and their status/position in a learning community.

SOCIAL NETWORK ANALYSIS IN MOOCS

Social Network Analysis (SNA) is a methodology that has become increasingly popular in the realm of education research (De Laat, Lally, Lipponen, & Simons. 2007). It provides the theoretical and methodological tools to understand activities and social processes in which learning occurs through extracting the patterns of connections among learners. In the simplest form, three points of data—two actors and the tie or link between them—comprise the basic unit of analysis. "Nodes" or "actors" are people, organizations, computers, or any other entity that processes or exchanges information or resources. While "Ties" or "edges" between nodes represent the information exchange, such as communication between a teacher and a student (De Laat et al., 2007). SNA draws the relationships among the actors and allows researchers to quantify the importance of the actors and detect clusters of actors that are more connected among each other than the random average (De Laat et al., 2007).

Another important concept is density, a network property that describes the general level of linkage among the nodes in an interaction network. The density of a network is defined as the number of ties in a network divided by the maximum number of all possible ties (Scott, 1991, pp. 72-73). Thus, the density of a network is at a maximum

when all the nodes are connected to each other. The density value of a network varies between 0 and 1. In a dense network (density score is close to 1), participants are more likely to mutually influence each other, and information tends to circulate in the network rapidly because participants are tightly interconnected (Hanneman & Riddle, 2005). Thus, a dense network can mediate peer interactions in a more efficient way. It offers more possibilities for participants to be engaged in the discussions of different subgroups, and to pursue pedagogically valuable discourse; it may also facilitate collective responsibility for advancing knowledge and distribution of expertise among participants.

While density describes the extent to which all participants are interconnected, modularity focuses on the structure of networks. It was designed to measure the strength of division of a network into groups (also called modules). Groups are divided based on the density of connections between the nodes. Nodes within the same group generally have denser connections than with nodes in other groups (Newman, 2006).

Besides density and modularity, centrality is another important parameter that reflects the behavior of individual participants within a network. It measures the extent to which an individual interacts with other individuals in the network. The more an individual connects to others in a network, the greater their centrality in the network (Wasserman & Faust, 1994). The following section discusses four common types of centrality measures that are relevant in this study.

Centrality measures in SNA

A common approach of SNA is structural network analysis, which mainly measures the actors' centrality such as degree, closeness, and betweenness. It is often used to study the relationship between an actor's position and his/her performance, persistence or the quality of contribution in a community (Van der Hulst, 2009).

Furthermore, it is used to identify influential actors and sub-communities. Specifically, the definitions and functions of degree, closeness, betweenness and Eigen centrality are provided in Table 2 (Van der Hulst, 2009):

Table 2: The Definitions and Functions of the Four Centrality Measures

Centrality measures	Definition	Function
Degree centrality	The number of links held by each node. In other words, how many direct, 'one hop' connections each node has to other nodes within the network. For a node in a directed network, in-degree refers to the number of edges coming towards the node, while out-degree refers to the number of edges going out from the node.	For finding highly connected individuals, popular individuals, individuals who are likely to hold most information or individuals who can quickly connect with the wider network. In educational settings, learners with higher degree are more active or popular learners who initiate or receive more interaction than others
Table 2, continued.		
Eigen centrality	A node's influence based on the number of links it has to other nodes within the network. Eigen	By evaluating the quality of the connections of a node (how well-connected one's

	centrality goes a step further than degree centrality by also taking into account how well connected a node is, and how many links their connections have through the network.	connections are), Eigen centrality can identify nodes with greater influence over the whole network. In educational settings, learners with higher Eigen centrality are those who initiate or receive interaction from a lot of well connected others.
Closeness centrality	The average distance between each node to all other nodes within the network. It calculates the shortest paths between all nodes, then assigns each node a score based on its sum of shortest paths.	For finding the individuals who are best placed to influence the entire network. In educational settings, learners with higher closeness centrality are those who have direct interaction with many other learners in the network

Table 2, continued.

Betweenness centrality	The number of times a node lies on the shortest path between other nodes. It shows which nodes act	For finding the individuals who influence the information flow within a network. In

	as ‘bridges’ between nodes in a network. It does this by identifying all the shortest paths and then counting how many times each node falls on the shortest paths.	educational settings, learners with higher betweenness centrality are those whose posts trigger a lot of discussions among other learners.

In summary, degree centrality is the simplest measure of node connectivity. It’s useful to look at in-degree (number of inbound links) and out-degree (number of outbound links) as distinct measures. Betweenness, on the other hand, is useful for analyzing communication dynamics. A high betweenness count could indicate someone holds authority over, or controls the communication between two other members in a network. By contrast, closeness centrality can help find good ‘broadcasters’, someone who can help spread information more quickly within the network due to their closer connection to every other member in the network. Eigen centrality is useful in identifying nodes that are not only well-connected to others, but also connected with more influential nodes within the network (Van der Hulst, 2009).

Using SNA to examine learner network in MOOCs

In the context of MOOCs, SNA is often used to untangle the complex learner network and to generate understanding about the characteristics of social networks in MOOC discussions and their implications for teaching and learning.

Passive and active participation in discussion forums

SNA can be used to identify passive and active participants in discussion forums. Passive participants, a term usually referred to as "lurkers", are often defined those who never post but read the group's postings regularly (Neelen & Fetter, 2010; Nonnecke, Andrews, & Preece, 2006). In online learning settings, it is well documented that passive participants are more prevalent than active participants who regularly contribute posts, and the majority of the content in an online community is created by the minority of the users (Arthur, 2006; van Mierlo, 2014). In reality, almost all participants, active or passive, read more posts than actually creating posts in discussion forums (Ebner, Holzinger, & Catarci, 2005). Even though lurkers comprise such a large proportion of website users, researchers have paid little attention to the lurking phenomenon until recent years. Researchers have identified multiple reasons for lurking, such as community culture, users' personality and the relationship between users and the group (Du, 2006; Fan, Wu, & Chiang, 2009; Nonnecke, 2000; Nonnecke & Preece, 2001; Tedjamulia, Dean, Olsen, & Albrecht, 2005). SNA is often used to make visible the passive participation structure in a learning community (Honeychurch, Bozkurt, Singh, & Koutropoulos, 2017). Examining the dynamics of passive participation is equally important with analyzing the active participation network, since lurking is a natural process for new comers to become active members of a learning community, or equally benefit from the knowledge exchange by observing, which can be perceived as legitimate peripheral participation by Lave and Wenger (1999).

The factors that influence social connections

One strand of the literature of social network studies in MOOCs explores the factors that influence social connections. For instance, Kellogg, Booth, and Oliver (2014)

studied the structure of peer support networks formed in two MOOCs designed for educators (on digital learning and mathematics learning) and examined factors that might account for such structures. The discussions in each course were studied as a single network separately. They found some cross-course consistency in general network measures and participation patterns: both networks had clear core-periphery structure and low edge weight, meaning that the interactions were relatively sparse among learners; also, learners' participation fell into four types of patterns, including mutual interactions, extensive but non-mutual interactions, thread initiation, and unresponded interaction attempts. They also found demographic factors affect network connections, though they were largely inconsistent across courses: the tendency that learners connect more with those who had the same activity pattern, taught at the same schooling level (elementary vs high school), or lived in the same state or country were only found in one MOOC. Joksimović et al. (2016) also investigated the factors that influence social connections. The study was conducted on two iterations of a programming MOOC, offered in English and Spanish respectively. A directed social network was extracted from the forums of the whole course. Similar to the results of the study of Kellogg et al. (2014), this research found some consistency of participation across the two MOOCs. However, they didn't find any cross-MOOC consistency in the association between social connection and learners' similarity in geographical location.

Centrality measures and learning outcomes

As mentioned above, centrality measures are useful in identifying learners that are active and influential in a learner network. Joksimović et al. (2016) examined the association between degree, closeness, betweenness centrality and academic performance (completion and distinction status). Results showed that degree centrality was

significantly associated with learning outcome across two MOOCs; effect of betweenness and closeness were only found in one MOOC but not in the other. Jiang et al., (2014) also examined associations between social centrality and academic performance (certificate, completion, and distinction status). They conducted the study on MOOCs in algebra and finance. Undirected social networks were extracted from the whole discussion forums based on co-presence in the same thread. The results found from the two courses were inconsistent: degree and betweenness were positively correlated with learning performance in the algebra course while no significant correlation was found between any centrality index and learning performance in the finance course. Contradicting to the findings of Joksimović et al. (2016), this study found learners tend to talk to those who are in different performance groups more than within the same group. In summary, the findings of these studies about social networks and learning are largely inconsistent or contradictory. One possible explanation is MOOC discussion forums are used for highly diversified purposes, such as understanding learning materials, clarifying course policy, and developing social connections. Consequently, analyzing the discussion forums as one social network may compile interactions with distinct natures together, confounding relationships and concealing important patterns (Wise, 2018).

Instead of only focusing on how centrality measures affect learning outcomes, some studies combined both centrality measures and the quantity of participation in the discussion forums to predict learning outcomes. For instance, Houston et al. (2017) built linear models to predict final course grades in a MOOC on innovation (two offerings) and a MOOC on programming. The prediction was based on the number of threads a learner contributed to, the degree, betweenness, closeness, and Bonacich power (a learner's connectedness to influential learners in the network). For each course, three social networks based on co-presence of learners in a common thread were built for three

weekly sub-forums selected from the beginning, middle, and end of the course. It was found that the predictors only correlated significantly with final course grade in some of the networks, with number of threads being the feature that most frequently correlated with course grade (6 out of 9 networks), followed by degree, betweenness, closeness, and Bonacich power; the correlations were all relatively weak ($r < .20$). Moreover, adding the centrality measures to linear regression models based on the number of threads contributed to did not explain significantly more variance in course grade. In addition to combining quantity of participation and network measures, some studies have also incorporated non-discussion indices of learning activity in the prediction of learning outcome. For instance, Guo and Wu (2015) used the quantity of discussion forums contribution together with learner's use of MOOC components to predict pass / fail for each unit. The quantity of forums contribution was measured by the number of forums *posts and total number of words posted by a learner; other predictors included* number of sessions a student logged into the course platform, page-load requests, total video playing time, and numbers of videos played, rewinds, pauses, fast forwards, and slow plays. They found that homework grade was significantly correlated with both forum variables: number of posts ($r = .20$) and total number of words ($r = .32$). Moreover, total number of words was found useful for predicting unit pass / fail status along with several of the non-forums variables. Besides the frequency of posting and the length of posts, earlier studies also indicate that the timing of posting plays a role in participants' reactions to forum posts (Jaech, Zayats, Fang, Ostendorf, & Hajishirzi, 2015; Lampe & Resnick, 2004; Lu & Farzan, 2015; Mazzolini & Maddison, 2007; Yusof & Rahman, 2009), although the findings are inconclusive. In another example, Jiang and her colleagues (2014) used forum and non-forum related indicators in week 1 of an introductory biology MOOC to predict the type of certificate learners obtained at the end

of the course. The predictors included learner's degree in a social network based on direct-reply relationship, average quiz score, number of peer assessments completed, and whether or not the learner was an incoming student of the university that offered the MOOC (who received incentives to participate). Two logistic regression models were built for predicting distinction vs. normal certificate and normal vs. no certificate. For the distinction vs. normal certificate model, degree in the learner network was found to be a significant predictor together with number of peer assessments completed and being an incoming student of the university. However, degree in the learner network was not a significant predictor in the normal vs. no certificate model. In summary, the inclusion of other course participation factors adds certain predicting power to students' learning outcomes, but what factors are the most significant predictors varied across different studies.

Network features and discussion topics

While MOOC forums often concern highly diversified topics and purposes, recently, a small number of studies have begun to differentiate analysis of interactions based on the topics of discussion. For example, Gillani and Eynon (2014) built networks based on co-presence within a thread in a business MOOC for each of seven sub-forums: Readings, Lectures, Cases, Final Project, Course Material Feedback, Technical Feedback, and Study Groups. Sub-forums networks consisted of largely distinct groups of learners and showed different levels of participant persistence over time, with "Cases" (for discussing learning material) showing the highest level of persistence overall. In another study, Gillani, Yasseri, Eynon, and Hjorth (2014) built social networks for two successive offerings of a business MOOC based on thread co-presence in eight sub-forums: Readings, Lectures, Cases, Final Project, Questions for Professor, Course Material

Feedback, Technical Feedback, and Study Groups. For both offerings, the proportion of one-off ties (edges with a weight of one) differed across sub-forums: highest for "Feedback" (used for technical support) and lowest for "Cases" (for discussing learning material). In addition, densities of interaction differed, with "Cases" and "Final Project" (both used for working on course material) showing greater cohesiveness than "Study Groups" (used for locating study partners). Although separating the network based on the pre-designated topic of sub-forums is straightforward and easy to operate, this approach comes with limitations. First, each MOOC sets different sub-forums, often defined quite narrowly in relation to the specific course, therefore the generalizability of findings based on such divisions is limited. Second, prior studies have shown that misplaced postings across sub-forums are very common in MOOCs (Rossi & Gnawali, 2014). As a result, social networks built based on sub-forums may not always accurately reflect the nature of relationships formed in forums interactions. An alternative approach differentiates interactions based on what is actually discussed by looking at the language used in the posts. This requires setting up a coding scheme for categorization and coding posts accordingly. The advantage of this approach is the ability to designate a useful set of groupings that is concise and generalizable (rather than tailored to a particular topic only for one MOOC) and more accurate categorization based on the actual posts themselves. Poquet and Dawson (2016) adopted such an approach to study a MOOC on natural science. They manually coded all discussions into five categories: "cognitive task" (conversations about material related to quizzes and assignment), "social task" (conversations about learner emotions about the assignments), "cognitive non-task" (conversations about the course topic not directly related to assignments), "social non-task" (purely social aspects), and "administrative/technical issues" (conversations about tools etc.). Constructing an undirected social network for the whole forums based on

thread co-presence, they found that these discussion topics did not significantly explain network formation. They suggested that this may be because different kinds of interactions play a role at different times in the course. In addition, it is possible that the categories used here were overly refined. Good learning discussion that facilitate the flow of useful information around the shared pursuit of knowledge may contain both comments about learning materials offered in MOOC, comments about the course topic not directly related to the materials in MOOC, as well as learners' emotion surrounding the topics being discussed (Wise, 2018).

Summary

The review of the literature reveals two critical issues that warrant attention. First, although the quantitative measures of forum participation (e.g., frequency of posting, length of posts etc.) and centrality parameters are often found to be positively associated with learning outcome, the correlations are relatively weak and the results from prediction studies are inconclusive as to what centrality parameters are consistently useful predictors of learning outcome. Conceptually, the quantitative measures of forum participation directly track the amount of participation while the network centrality parameters indicate the social relationships among learners (Wise, Cui, & Jin, 2017). These measures reflect distinct aspects of forum participation, therefore examining and comparing their usefulness for predicting learning outcome can contribute to our understanding of student learning in MOOC forums. Second, little prior research has taken into consideration the content of discussions when examining the association between forum participation and learning outcome. Discussions in MOOC forums often involve a broad range of topics that may or may not relate to learning the course content (Stump, DeBoer, Whittinghill, & Breslow, 2013) and thus would be more or less

expected to predict learning outcome. There is a need to examine the content of learners' discussions to see whether the qualitative nature of their posts, in combination with the quantitative measures, predict their learning outcome in a more accurate way.

CONTENT ANALYSIS ON DISCUSSION FORUMS IN MOOCS

Across academic fields, there has been a burgeoning literature demonstrating the usefulness of language and discourse in predicting a number of psychological, affective, cognitive, and social phenomena, ranging from personality to emotion to learning to successful group interactions (Chung & Pennebaker, 2014). Within the educational contexts, there are many critical learning-related constructs that cannot be directly measured, but can be inferred from measurable signals like language and other behavioral patterns. Working with these barriers, researchers are continually pushing beyond the boundaries of established approach of analytics. In this realm, it is particularly important that these endeavors are guided by established theories. A number of psychological models of discourse comprehension and learning, such as the construction integration, constructionist, and indexical-embodiment models, are often used to the exploration of learning related phenomena in computer-mediated educational environments (Dowell et al., 2015). These psychological frameworks have identified the representations, structures, strategies, and processes at multiple levels of discourse.

With regard to analytical approaches, there has been extensive knowledge obtained from manual content analyses of learners' discourse during educational interactions, however, these methods are no longer a viable option when the number of learners reach a certain magnitude. Therefore, researchers have been incorporating automated linguistic analysis, including more shallow level word counts and deeper level discourse analysis approaches. Both levels of linguistic analysis are informative. Content

analysis using word-counting methods enables a quick overview of learners' participation levels, as well as assessing specific words. For instance, a study by Wen, Yang and Rose (2014) is an example of incorporating word counts of theory-informed and carefully selected words with manual text coding. Their work tested the correlation between specific words used by the students and the degree of their engagement and commitment to complete the MOOC. To extend analysis of text beyond the shallow level word counts, researchers began to conduct a deeper level discourse analysis employing sophisticated natural language processing (NLP) techniques (e.g., syntactic parsing and cohesion computation). A thorough literature review reveals the following seven themes of inquiry regarding content analysis in discussion forums in MOOCs: identifying topics, examining cognitive process, analyzing emotions/sentiments, identifying content-related posts, detecting confusion, detecting help-seeking posts, revealing linguistic features.

Identifying topics

Identifying topics is one of the most common purposes to use NLP to analyze discussion forums content, in order to assist MOOC instructors to efficiently navigate through the massive amount of content generated by learners in MOOCs. Among all topic-modeling techniques, Latent Dirichlet Allocation (LDA) is the most popular method. Atapattu and Falkner (2016) proposed a model based on LDA to facilitate educators to locate relevant content more effectively. Similarly, Wong, Wong and Hindle (2019) used both supervised and unsupervised LDA methods to extract topics from forums posts. Other researchers used more sophisticated LDA techniques to classify words based on the different topics in a MOOC. For example, Ramesh and his colleagues (2014) used seeded LDA to point words to specific types of topics according to the syllabus of the course, which seems to be more informative than merely differentiating

content-related posts and those dealing with the logistics of course (e.g., the location of certain resource, the due date of assignments etc.). Besides assigning topics based on course topics, Rossi and Gnawali (2014) proposed a different approach of identifying topics based on the components of the course, such as assignments and lectures. Their classification model achieved a satisfactory performance. However, although LDA provides a quick review of topics discussed in MOOCs, the classifications are often ambiguous and coarse-grain, with many chaotic posts randomly assigned to unrelated categories, which makes it difficult for human to interpret the results and generate little insights for instructors to facilitate the discussions among learners.

Examining cognitive process

Researchers also use NLP to analyze cognitive process in MOOC to better understand students' behaviors. For example, Wong, Pursel, Divinsky, and Jansen (2015) built a model to analyze topics and words and categorize posts according to the six levels of Bloom's Taxonomy. The results showed that students in high-level cognitive process made faster progress in learning. Moore, Oliver and Wang (2019) used a similar method to identify the factors that predict cognitive process in learning in MOOCs. They found that there was a significant correlation between the number of words and cognitive processes, but there was no correlation between students' logical thinking, educational background and their participation in learning. Moreover, they found that the factors predicting students' cognitive process were not only limited in participation in discussion forums, but also the duration of watching instructional videos, completing reading materials and so on.

Analyzing emotions/sentiments

The content of the discussion forums is also often analyzed with the purpose to detect learners' sentiments and emotions, and to examine the relationship between sentiments/emotions and performance. For example, Wen et al. (2014) employed sentiment analysis to mine the emotions in forums posts to better understand the students' struggles in participation and the reasons for their withdrawal. In order to better study whether students would be affected by the emotions expressed in other peers' posts, the researchers used a survival model to measure students' emotional changes. The results showed that there was a significant relationship between the dropout rate and learners' sentiments towards the course. If students express negative emotions, they are more likely to drop out. Hu, Dowell, Brooks and Yan (2018) used LIWC, an automatic text analysis tool, to further study the affiliation and emotions of learners in MOOC and their development over time. Their method classified words into psychological categories, such as "nice" and "hope" for positive emotions, while "nasty" and "hate" for negative emotions and "friend" for affiliation. Through the study of five different types of MOOC discussion forums, they found that, with the passage of time, students' negative emotions towards the course increased, while affiliation decreased, but positive emotions did not change significantly. As the course proceeded, students' negative emotions toward the course exceeded positive emotions. This explained why students' performance was below expectation and the drop-out rate was high in that MOOC. Despite all these research efforts, sentiment detection is challenging for two reasons. First, from the technical perspective, most of the current studies only detect sentiments/emotions based on the occurrences of certain keywords that signal sentiments/emotions, while disregarding the actual topics that give rise to those sentiments/emotions. In other words, the algorithms are not capable of identifying the actual topics that trigger those sentiments/emotions,

while many of those topics may be completely unrelated to the learning design of the course, or students' learning experience. For example, a learner may express anger towards a personal experience in the past, which is not related to the his or her own learning experience in the course. Linking these sentiments/emotions to students' learning outcome will result in misleading conclusions that bear limited useful implications for instructional design in MOOCs. Second, from the perspective of sample selection, all content analysis research only considers learners with text input in the forums, while excluding a large number of "lurkers" who access course materials but remain silent throughout the course. Consequently, the findings may be biased to a great extent.

Identifying content-related posts

In order to help MOOC instructors and facilitators efficiently navigate through the huge amount of text in discussion forums, many scholars developed models to automatically identify content that are relevant to the course topics and filter irrelevant information. For example, Wise et al. (2016) built a linguistic analysis model based on manually coded and validated discussion forum data to distinguish between content-related and content-unrelated posts. They defined content-related posts as those that seek help or give relevant information and comments or share relevant course content. They tested the model on content-related posts, which are important sources for us to understand learners' cognitive process. Some researchers took further steps to analyze these content-related posts trying to identify the types of discussions that are more conducive to learning. For example, Wang and his colleagues (2015) established coding algorithm based on Chi and Wylie's (2009) framework for differentiating learning activities. The purpose is to study the relationship between students' cognitive behavior,

quality and quantity of participation and their learning outcomes. They divided the forums posts into three categories: active discourse, constructive discourse and interactive discourse. They found that students' active and constructive discussion behaviors significantly predict their learning gains. While interactive discussion behaviors are significant in predicting learning gains only for students who are less active in the forums.

Detecting confusion

Detecting confusion is another important task for text analysis in discussion forums, so as to locate learners who need help and find out what kind of assistance they need. Agrawal et al. (2015) developed an educational support tool YouEDU using the data from the Stanford MOOC platform. Their tool aims to automatically discover and locate posts that signal confusion in discussion forums. Firstly, students' emotions, judgments of emergencies and other descriptive variables were used as the basis for training. Secondly, for those who voiced confusion in the forums, their tool recommended relevant videos from the course that may potentially help them solve their problems. Finally, they used different MOOCs to validate the performance of the model and ranking algorithms to evaluate the relevance of recommended videos. The results show that the model achieved an average precision rate at 0.74. However, there are two problems with this approach. First, when students encounter the same confusion repeatedly, the algorithm will probably recommend the same video, which can cause negative emotions if the video is not helpful. Secondly, it does not take into account the evolutionary needs of students. Limited video resources cannot solve all the puzzles.

Detecting help-seeking posts

Detecting help-seeking posts is another common use of NLP to address students' urgent needs. Almatrafi, Johri and Rangwala (2018) established a sustainable and responsible classification model for various courses to locate urgent posts by investigating different semantic features and data mining classification algorithms. Hecking, Hoppe and Harrer (2015) also used the classification method to identify posts, with an average precision rate of 0.79. But these two studies only built their models based on a small number of MOOCs, therefore the model may not be applicable to courses of other domains. Chandrasekaran, Kan, Tan, and Ragupathi, (2015) used 61 courses to validate the performance of their classification model, but the accuracy rate was not ideal. This suggest that the model was not generalizable across MOOCs of different domains, which is a common challenge of large scale text analysis in online learning.

Revealing linguistic features

Content analysis is also useful for revealing certain linguistic features of learners' discourse to discern factors that contribute or hinder learning. For instance, Moon, Potdar and Martin (2014) attempted to identify student leaders solely based on textual features, or specifically by analyzing how student leaders influence other students' language use. They proposed an improved method of measuring language accommodation based on learners' choice of words given a semantic topic of interest, and showed that students with high degree centrality (defined as "student leaders") indeed coordinated other students' language usage. Their method also successfully distinguished student leaders from two MOOC with different topics. Dowell and colleagues (2014) explored the possibility of using discourse features to predict student performance during collaborative learning interactions. Their results indicated that students who engaged in deeper

cohesive integration and generated more complicated syntactic structures in their sentences performed significantly better. In line with this, another study by Dowell, Graesser, Tausczik and Pennebaker (2014) demonstrated that cognitive linguistic cues can be used in detecting students' socio-affective attitudes towards fellow students in online collaborative learning environments. As a whole, these studies highlight the critical and complex role of language and discourse. This is not surprising, since language is a primary means for expressing and communicating information in computer mediated learning environments.

Summary

With the trend of research interests shifting from quantitative analysis to qualitative analysis that provides more contextual information to understand student learning in MOOCs, the number of studies of content analysis in MOOC forums has been increasing rapidly over the past five years. However, early exploration of discussion forum data from different angles, as presented above, still faces many challenges: the level of analysis is still coarse grained, the content analysis models are preliminary and lack accuracy, the results are often ambiguous and not generalizable to MOOCs of diverse topics. These limitations need to be addressed by developing more robust text analysis models using the latest NLP techniques, and examine the forum posts based on more sophisticated theoretical frameworks that could potentially yield more useful insights to inform the design and facilitation in MOOCs.

AUTOMATIC TEXT ANALYSIS WITH DEEP LEARNING

In recent years, deep learning is rapidly gaining popularity in the field of computer science. It allows computational models that are composed of multiple

processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in the fields of speech recognition, visual recognition, object detection and many other domains (LeCun, Bengio, & Hinton, 2015). In the field of NLP, algorithms of deep learning have also been widely used. Among these algorithms, BERT (Bidirectional Encoder Representations from Transformers) stood out with exceptional performance ever since it was introduced by researchers at Google AI in 2018 (Devlin, Chang, Lee, & Toutanova, 2018). It was considered revolutionary in the Machine Learning community by excelling in a wide variety of NLP tasks due to its outstanding ability of bidirectional contextual learning of words. Prior to BERT, research has been extensively conducted to improve the performance for word embedding (Howard & Ruder, 2018; Pennington, Socher, & Manning, 2014; Peters et al., 2018; Rong, 2014), which is a technique that encodes words with numbers such that words often used in similar contexts would have similar numeric representation. Word embedding yields great improvements over the traditional bag of word (BOW) method which simply represents words with their frequency of occurrence in a sentence (Lilleberg, Zhu, & Zhang, 2015). However, most of the works on word embeddings fell short in understanding understand the contextual meanings of words. For example, previous word embeddings could not differentiate between the "bank" to withdraw money versus a river "bank" for a walk. Although recent works such as Embeddings from Language Models (ELMo) were able to learn language contexts (Peters et al., 2018), the nature of learning in such word embeddings was unidirectional or shallowly bidirectional, which was suboptimal for performance and harmful for transfer learning (Devlin et al., 2018). BERT, on the other hand, was able to learn better language representations through its original self-supervised learning methods such as masked language models and next sentence prediction (Devlin et al., 2018). What is

more, BERT's use of bidirectional self-attention allows it to incorporate a greater context when encoding sentences rather than just context from previous words (Devlin et al., 2018).

Besides the advantage of bidirectional contextual learning, BERT's superior performance is also due to leveraging the benefits of its pre-trained parameters for transfer learning. Transfer learning allows models to inherit an outstanding base performance on tasks such as natural language inference and paraphrasing (Fedus, Goodfellow, & Dai, 2018), which helped generate more accurate results for tasks such as text classification (Liu, Sun, Lin, & Wang, 2016). Recent works have shown that BERT for transfer learning can significantly improve model performance even with limited data (Lan et al., 2019). For example, Liu et al. (2019) achieved an accuracy of 83.2% on a dataset for high school English reading comprehension by optimizing BERT, while the previous best prediction accuracy rate based on the same dataset was 44.1% in 2017. Nogueira and Cho (2019) adopted BERT to attend a data mining competition for question-answering. Their results ranked the 1st among all the contestants, and have outperformed the previous best score by 27%. Therefore, in this study, BERT was selected for text analysis.

SUMMARY OF LITERATURE

The review of literature reveals the following existing gaps in empirical studies: 1) The current studies that examined social presence are limited in bounded online learning contexts (e.g., synchronous small-scale LMS courses) instead of MOOCs; 2) The current analyses of social presence are not in relation to learners' level of engagement in the learner network in MOOCs, namely their status or positions in the network; 3) The existing text analysis methods are based on traditional machine learning

algorithms and have relatively low accuracy of classification, especially when it comes to classifying text into a complex framework with more categories.

Given the complexity of the interaction patterns in MOOCs, as well as the diverse demographic backgrounds of learners, the role of social presence in this particular context remains under-explored and requires further clarification. To this end, this study specifically investigates one MOOC to gain an in-depth understanding of learners' interactions from the perspective of social presence.

Chapter 3: Methodology

This chapter outlines the study design, introduces the research questions, presents the context of research, data collection instruments and procedures, and finally discusses the approaches to carry out the desired analyses.

STUDY DESIGN

In order to answer the research questions in this study, a mixed method approach was employed (Tashakkori & Teddlie, 2003). This approach is informed by pragmatism from an epistemological sense. Pragmatism is a deconstructive paradigm that advocates the use of mixed methods in research, "sidesteps the contentious issues of truth and reality" (Feilzer, 2010, p. 8), and "focuses instead on 'what works' as the truth regarding

the research questions under investigation" (Tashakkori & Teddlie 2003, p. 713). The mixed method approach consists of procedures to collect, analyze and integrate both quantitative and qualitative data in different stages of the research process in the study (Creswell, 2005). In this study, neither quantitative nor qualitative methods could adequately within themselves cover the scopes and depths of the research questions pertaining to learners' participation patterns, social presence and learning outcomes in the MOOC. When combining the two, quantitative and qualitative methods complement each other and provide a holistic and in-depth view of the research problem (Green, Caracelli, & Graham, 1989; Tashakkori, Teddlie, & Teddlie, 1998).

Specifically, the qualitative data came from learners' posts in the six discussion forums in the MOOC; while the quantitative data included the system log which recorded learners' posting behaviors (e.g., number of posts, time of posting etc.), and the survey responses by the end of the MOOC regarding learners' certificate status, perceived learning and satisfaction.

The collected data were analyzed from both qualitative and quantitative aspects: 1) To qualitatively analyze the forum posts, a text classification model using the most up-to-date computational algorithms in NLP, built and validated based on the previous offering of the same MOOC, was adopted to automatically classify discussion posts into 13 different categories of social presence; 2) To quantitative analyze the data, four SNA parameters were used to measure learners' network centrality (namely in-degree, closeness, betweenness and Eigen centrality). Then, correlation and regression analyses were conducted to discern the relationships between social presence and learners' network centrality. Learners' posting behaviors (e.g., frequency of posting, average length of posts, day of posting) were also included in the regression models to see whether they add any predictive power to learners' centrality. Finally, correlation

analyses were conducted to see how learners' network status correlate with their learning outcome. The purpose of using mixed methods is to see whether the qualitative nature of the posts, in combination with learners' posting behaviors, predict learners' position in the learning community, and how learners' network status correlate with their learning outcome.

RESEARCH QUESTIONS

In an attempt to better understand learners' participation patterns and social presence in the discussion forums in a MOOC, this study attempts to answer the following research questions:

1. What social presence did learners exhibit in the discussion forums of the MOOC?
2. How does the structure of learner network change over the six modules of the MOOC? And how does the learners' social presence differ as the learner network evolves over time?
3. What is the relationship between learners' social presence and their centrality in the learner network? And how do learners' posting behaviors contribute to the prediction of their centrality?
4. How do learners' network centrality correlate with their learning outcomes (measured by certificate status, perceived learning and satisfaction)?

RESEARCH CONTEXT

The data source for this study came from a MOOC titled *Data Journalism and Visualization with Free Tools*, developed and maintained by the University of Texas at Austin. This MOOC consists of six modules and lasted around four months from mid-

October 2019 to mid-February 2020. The components of this MOOC include instructional videos, reading, quizzes and discussion forums. For each module, learners were required to watch the videos and finish the assigned reading before they access the forums to answer and discuss open-ended questions with their peers. Typically, the instructor assigned one or two tasks in each forum and asked learners to complete the tasks/assignments using the concepts and skills explained in the videos and reading of the that module. Each learner would have to start a new discussion thread to post their responses to the assignment (see Figure 2). After posting, they were encouraged to read their peers' posts and give comments and suggestions. Commenting on peers' posts was required in certain modules but optional in others. All comments to a students' assignment response were nested under the discussion thread started by that student (see Figure 3). In other words, each discussion thread starts with a student's response to the forum tasks/assignments, which is followed by posts of comments from his or her peers. The use of discussion forums in this MOOC provides a space for active interaction among learners to deepen the understanding of the course topic and co-construct knowledge by stimulating critical discourse. To obtain the certificate by the end of the MOOC, learners were required to provide their solutions to the tasks that the instructor posted in the forums. Table 3 outlines the specific tasks and requirements in each module.

Table 3: The Topics, Tasks and Requirements in the Forum of Each Module

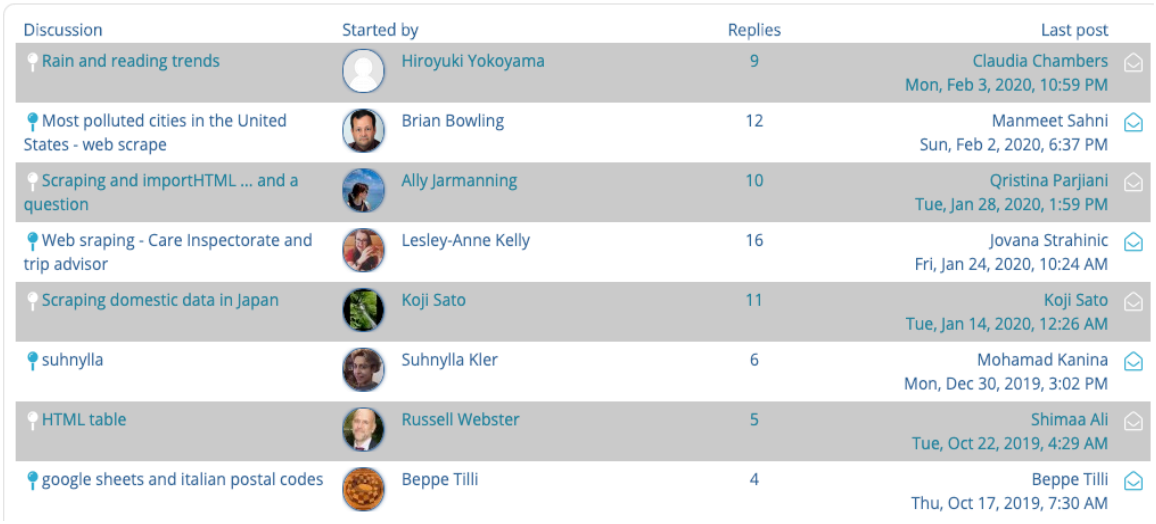
Module	Topic	Tasks and Requirements
1	Finding and getting data	1. Find data about a topic you're interested in. Send a brief message to the forums explaining what the data is about, why it matters, and how you obtained it. If possible, suggest what stories or visualizations you envision as doable based on the data. Once you've sent your own message to the forums, read messages from other students. **Reply to at least TWO of them offering constructive comments, feedback, or even suggestions.** 2. Find two sites that are different from those featured in Marco's import HTML & Web Scraper video classes. The first site must contain an HTML table.
2	Preparing data	**1.** Clean up a data set of your choice based on what you've learned this week. It could be a data set about a topic you like, or a data set mentioned in the video lectures. Send a short post to the discussion forums explaining very briefly what you did and what problems you found. Once you've sent your own message to the forums, read messages from other students. **Reply to at least TWO of them offering constructive comments, feedback, or even suggestions.** **2.** What are the differences between tables that were made for humans and tables that were made for computers? What are the situations where one might be better than the other? Explain in your words what makes a table "clean" and **evaluate a colleague's response.**

Table 3, continued.

3	Finding stories in data	1. Explore the winning dashboards from the Visualize 2030 contest. Identify the visualization that you think tells the most compelling story. List 3 reasons why you find that visualization interesting and 1 improvement that you would make. 2. Identify a chart type that is not available in Data Studio. Make a post in the forum naming the chart type, link to at least one example visualization that uses that chart type, and explain specific use-cases. 3. (Optional/bonus question) Identify an infographic or visualization from the MakeoverMonday Series. The underlying data for each of the viz is available at the site. Try to replicate the one visualization or at least a part of it in Data Studio while making your own improvements. Share your finished Data Studio dashboard link and comment on what improvements you made and why.
4	Machine learning in data journalism	1. How might you use Machine Learning to report a story? List 2-3 examples. 2. Machine Learning models are sometimes described as "black boxes." We can't always tell why they make the decisions they do. Is this a problem for data journalism? Can we work around it?

Table 3, continued.

5	Visualizing data	1. Share any visualization you like (or dislike) in the forum. Explain why you chose it, using the language and concepts you've learned in this module. After that, **reply to the choices of at least TWO other students.** Be constructive in your comments! 2. Design at least one visualization with Flourish (you can design more if you want). You can use any data set you want. It can be data we've used in this MOOC so far, or any other you've found interesting. Publish your visualization, and share the link in the forums. After that, **gently send feedback to at least TWO other students** on the visualizations they designed.
6	Data-driven storytelling	1. What has been the most memorable data story that you've ever seen? Why do you think it stuck with you? Please share a link if possible. After that, **comment on the choices of at least TWO other students**. Be constructive in your responses! **2.** What is a question you have that you think would make for a good data story and why? Based on it, try to design your own data story—or at least part of it—using one of the data sets we've used so far in this course, or any other data set you wish. Publish it and share it in the forums. After that, **send constructive feedback to at least TWO other students.**



Discussion	Started by	Replies	Last post
Rain and reading trends	Hiroyuki Yokoyama	9	Claudia Chambers Mon, Feb 3, 2020, 10:59 PM
Most polluted cities in the United States - web scrape	Brian Bowling	12	Manmeet Sahni Sun, Feb 2, 2020, 6:37 PM
Scraping and importHTML ... and a question	Ally Jarmanning	10	Qristina Parjiani Tue, Jan 28, 2020, 1:59 PM
Web sraping - Care Inspectorate and trip advisor	Lesley-Anne Kelly	16	Jovana Strahinic Fri, Jan 24, 2020, 10:24 AM
Scraping domestic data in Japan	Koji Sato	11	Koji Sato Tue, Jan 14, 2020, 12:26 AM
suhnylla	Suhnylla Kler	6	Mohamad Kanina Mon, Dec 30, 2019, 3:02 PM
HTML table	Russell Webster	5	Shimaa Ali Tue, Oct 22, 2019, 4:29 AM
google sheets and italian postal codes	Beppe Tilli	4	Beppe Tilli Thu, Oct 17, 2019, 7:30 AM

Figure 2: An Example of the Discussion Threads within One Forum

Figure 3: An Example of the Posts within One Discussion Thread

PARTICIPANTS

The total enrollment of the MOOC is 6417 students who came from 162 countries. The top five countries that had the largest enrollment were the United States, India, Brazil, the United Kingdom and Spain. However, given the purpose of this study, only participants of the discussion forums were selected (n = 2479). The criterion of selecting participants is any individual who had read or contribute at least one post in any module of the MOOC. Among these forum participants, 721 of them had contributed at least one post in the forum in any module. A survey was distributed to these post creators (n = 721) by the end of the MOOC to ask about their perceived learning and satisfaction. The demographic information of the survey respondents (n = 71) is shown in Table 4.

Table 4: Demographic Information of the Survey Respondents (n = 71)

	Number of participants	Percentage of participants
Gender		
Male	45	63%
Female	26	37%
Ethnicity		
Asian or Pacific Islander	9	13%
Black/African American	5	7%
Hispanic or Latino	10	14%
Mixed race and others	7	10%
White/Caucasian	40	56%

Table 4, continued.

Education		
High school graduate	3	4%
Associate degree in college (2-year)	3	4%
Bachelor's degree in college (4-year)	21	30%
Master's degree	30	42%
Doctoral degree	5	7%
Other degree (e.g., MD, JD)	9	13%
English Proficiency		
Native	23	32%
intermediate	16	23%
Advanced	32	45%

DATA SOURCE

To answer the research questions, three types of data were collected: 1) learners' posts in the forums over the six modules; 2) the system log data that record the details of each post; and 3) the survey responses from post creators about their perceived learning and satisfaction.

Forum posts

A total of 8381 of learners' posts over the six modules were extracted for analysis. As shown in Figure 2 and Figure 3, by the end of each module, the instructor created a forum and posted questions for discussion regarding the topic of that module. Each student was expected to start a new discussion thread to post his answers to the questions in that forum. After posting their own responses to the forum assignment,

students were encouraged to enter their peers' threads to leave comments or feedback. In other words, each discussion thread was created by one individual and began with his or her own response to the tasks in the forum, and all comments towards his or her response were nested under the same discussion thread.

System log data

The main purpose of collecting the system log data is to obtain the details of each post to learn more about the posting behavior of each individual, which may also contribute to one's status in a learner network. The possession of these data enables the calculation of: 1) the frequency of reading posts and creating posts of each forum participant; 2) the timing of their posting behaviors, which is reflected in the time difference between the publishing time of a specific post and the very first post in that particular forum. This information reflects how early a learner contributes his or her post in a forum; 3) the length of each post. The number of words of each post is also counted for the sake of measuring the length of each post.

Survey

A survey was distributed by the end of the MOOC asking forum participants regarding their perceived learning and satisfaction (Table 5). The items regarding students' perceived learning were adapted from the learning perception survey devised by Watson, Kim, & Watson (2016), which was grounded in the literature about students' perception of learning and attitude change (Kamradt & Kamradt, 1999; Simonson, 1979; Simonson & Maushak, 1996; Scott & Wheeles, 1977), and also used in MOOC settings. The items of this survey seek to address four areas of learning: *General learning, Cognitive learning, Affective learning,* and *Behavioral learning*. The *General learning*

items examine learners' internalized positive attitudes and perceptions toward content and their positive belief in accomplishing a learning task. The *Affective learning* items describe learners' emotions or how they feel towards the instructional topics. The *Cognitive learning* items capture the extent to which learners understand the topic. While *Behavioral learning* items concern learners' plans on change of actions after they finish the course. The survey items were measured on a 5-point Likert scale, scored from 1 (strongly disagree) to 5 (strongly agree). Participants were asked to rate to what extent he or she agrees with each item under the four dimensions of learning. The reliability with Cronbach's α values of the general learning, cognitive learning, affective learning, and behavior learning scales was .67, .68, .63, and .76 respectively.

The items regarding students' satisfaction were based on Strong, Irby, Wynn and McClure's (2012) satisfaction scale that was used to ascertain graduate students' satisfaction in distance courses (Table 5). Similarly, a 5-point Likert scale scored from 1 (strongly disagree) to 5 (strongly agree) was used for participants to rate to what extent he or she agrees with each item. The Cronbach's Alpha of the satisfaction items was .87. The reliability of the combined survey instrument was calculated *ex post facto* α = .88 resulting in a high degree of internal consistency (Cronbach, 1951).

Table 5: Survey Items that Measure Learners' Perceived Learning and Satisfaction

General Learning
1) I enjoyed the MOOC
2) I found the work within the MOOC exciting/stimulating and engaging
3) My perspective toward this topic has changed as a result of this MOOC
Cognitive Learning
1) I am more informed and knowledgeable about this topic after this MOOC

Table 5, continued.

2) I am more inclined to consider multiple perspectives about this topic after this MOOC
3) I agree with the perspective presented by the MOOC about this topic
Affective Learning
1) My feelings about this topic have changed as result of this MOOC
2) I feel more connected to this topic as result of this MOOC
3) I feel confident that my opinion about this topic is an informed and correct one
Behavioral Learning
1) I have specific plans for changing my lifestyle in regard to this topic
2) I plan on convincing others to make lifestyle changes regarding this topic
Satisfaction
1) I am satisfied with this MOOC
2) Participating in this MOOC is worth my time
3) I enjoy studying this MOOC
4) This MOOC is stimulating
5) This MOOC is exciting
6) I look forward to participating in another MOOC like this
7) I prefer learning in MOOCs

DATA COLLECTION PROCEDURE

Collecting survey data from the MOOC

The survey was built on Qualtrics and was distributed to the target participants, those who have contributed at least one post in any forum in the MOOC, at the end of the MOOC (mid-February 2020). A recruitment email was sent to all potential participants by the MOOC coordinator with a description to inform students about the purpose of this

study and the risk of participation. Participation is voluntary. The survey responses were stored on Qualtrics and later exported for analysis. A total of 84 responses were received. After removing 13 incomplete responses, 71 valid responses were exported for data analysis.

Collecting system log data from the MOOC

Upon the approval of the university, the log data that recorded learners' behaviors on the MOOC platform were downloaded from the system two weeks after the MOOC ended. These data inform the researcher about the participants' posting behaviors in the forums and their completion status, which is whether they received the course certificate. All data were de-identified by replacing the real name of each participant with an alias.

DATA ANALYSIS

Building a text classifier to identifying social presence in the posts

To answer research question 1, which investigates the distribution of social presence across the six modules, it is necessary to build a text classifier to identify different social presence indicators in the posts. Prior research regarding content analysis in MOOC forums is mostly based on manually coding a small set of learner-generated posts (Barak, Watted, & Haick, 2016; Kop, 2011). However, this type of qualitative analysis is time-consuming and labor intensive, thus not practical to deal with large scale text data. To combat this methodological challenge, data from the previous offering of the same MOOC were used to build a text analysis model in order to automate the classification process based on a well-validated social presence framework from Shea et al. (2010). It hypothesizes three dimensions of social presence: affective expression, open communication, and group cohesion, which are necessary to establish a sense of trust

and, ideally, membership in a community dedicated to joint knowledge construction. Under these three dimensions, there are multiple indicators that further describe each dimension of social presence. During the coding process, several social presence indicators were removed from the original framework (Table 1), such as *Addresses the group using inclusive pronouns*, *Use of humor* and *Unconventional emotion expression*. They were excluded because their occurrence was extremely rare (N < 10) in the dataset of this study, and such low occurrence could not make a reliable training dataset to train the text classifier. Table 6 shows the social presence indicators that were retained for training the text classifier.

There were three phases in building and validating the text classification model: 1) 3500 sentences from the discussion forums in the previous offering of the MOOC were randomly selected and manually coded into different social presence indicators (Table 6). Normally, a post often includes multiple sentences. In order to accurately capture each type of social presence and conduct more fine-grained analyses, each post was split into individual sentences based on the use of certain punctuations, namely, period, question mark and exclamation mark. Each sentence was assigned a code of social presence in the coding scheme. Three researchers were involved in this qualitative coding process. Each of them was randomly assigned around 1200 sentences for manual coding. They met constantly and compared codes until the agreement rate reached 100%. (2) All labeled data were used to train and validate the text classification model by testing different algorithms. (3) The best performing algorithm from phase 2 was applied to automatically analyze the posts at sentence-level into different social presence indicators.

Among the 13 social presence indicators in the final coding scheme (Table 5), the machine learning approach was only applied to identify ten of them. Three social presence indicators, namely *Asking questions, Sharing resources* and *Using Vocatives,*

were identified by detecting the presence of certain linguistic markers. Specifically, a sentence would be labeled as *Asking questions* if there is a question mark at the end of the sentence; if there is a hyperlink within the sentence which indicates the sharing of content on an external webpage, it would be labeled as *Sharing resources*; *Using Vocatives*, defined as addressing or referring to others by name, was identified based on the presence of a name in posts that matched any name within the students' name list in this MOOC. This simpler approach was adopted to capture these three social presence indicators because it is more efficient and effective than using machine learning approach.

For some specific cases, there are overlapping codes for an individual sentence. In other words, multiple codes were assigned to capture each social presence indicators in one sentence. For example, the sentence "Thanks for your detailed feedback, Shanker" will simultaneously be assigned two codes: *Expressing gratitude* and *Using vocatives*, since this sentence contains a thankful note and the name of a specific individual.

Aspects	Categories	Linguistic Markers	Examples
Affective expression	Positive emotions	Keywords such as "happy", "enjoy", "amazing", "learn a lot" etc.	i think the information you gathered is really interesting
	Negative emotions	Keywords such as "sad", "not happy", "disappointed", "frustrated" etc.	I got a bit confused now because i was later working with a lot of similar ones
	Self-disclosure	keywords such as "I'm from…", "I work in…", "I'm a fan of …" etc.	i'm also from ohio (dayton), so i get your hot water situation

Open Communication	Referencing others	Using the symbol "@" to reference a person; using keywords such as "like …(somebody) suggested", "comments by … (somebody)" etc.	Like @Becky suggested with 11 countries listed it could well be displayed on a world map
	Asking questions	Using the question mark	What do you think ? Improvements to suggest ?
	Complementing others	Using keywords such as "did a good job", "brilliant", "good idea" etc.	i find it very informative, also you did a great job with the legends and notes attached
	Expressing gratitude	Using keywords such as "thank you", "thanks", "grateful" etc.	Thanks, your comments are valid and important
	Expressing agreement	Using keywords such as "you are right", "I agree", etc.	i think you're right, it seems to describe the relationship better
	Disagreement/ doubts/criticism	Using keywords such as "I don't think…", "I disagree", "…something wrong", "I doubt…", "not a good idea/way/solution" etc.	i don't think map chart can add any value for this life expectancy information
	Offering advice	Using keywords such as "I suggest", "I advise", "why not…", "how about…", "may be better to…" etc.	i would suggest creating three charts for each of the year
	Personal opinion/reflection	Using keywords such as "I think…", "in my opinion", "I tried…", "my method is…" etc.	my trick was to use the "who won where" map three times to make comparisons easier

	Sharing resources	Giving the url to link to external resources	You can check out some examples here: https://www.tableau.com/learn/articles/best-beautiful-data-visualization-examples
Group Cohesion	Vocatives (addressing an individual by name)	Mentioning the name of a peer	Hi Chris, I went to the link that you included in your critique

Table 6: The Coding Scheme of Social Presence

In order to build an accurate text classifier, three language features were included in the candidate models: Part-of-Speech (POS), Named Entity Recognition (NER), and BERT's word embedding (WE) features (see the description of each in Table 7). Specifically, NER was considered an important feature because of the context of this MOOC. There were a lot of tasks in the forums where learners were required to create data visualization artifacts (e.g., interactive graphs/maps) of their chosen topics. For example, mortality rate of Ebola in Africa, suicide rate in Singapore, consumer satisfaction in France etc. Given the diverse backgrounds of learners, they chose vastly different topics. The descriptions of their topics involved a lot of specific organizations, locations, numbers, and duration. Raw text inputs concatenated with POS and NER tags were used to train the model in order to test if these linguistic features would add any value to the overall performance than training only raw text.

Table 7: Description of Features in Building the Text Classifier

Features	Description
Part-of-Speech Features (POS)	Grammar tag of each word such as verb, adjectives, and adverbs, punctuation
Named Entity Recognition Features (NER)	Information extracted from words such as names of people, organizations, and locations etc.
BERT's word embedding features (WE)	Numeric representations of words by using BERT's word embeddings (e.g., the word "book" and the word "dog" may have very different numeric vector representations that are distant from one another, since they are usually used in different contexts; but "cat" and "dog" may have numeric vector representations that are closer to one another)

To compare the performance of BERT with tradition machine learning algorithms, Random Forest was used to build a classifier to provide a base performance. Random Forest was selected because of its ensemble feature from a set of decision trees that usually yields desirable results (Xu, Guo, Ye, & Cheng, 2012; Kovanović et al., 2016; Liu, Kidziński, & Dillenbourg, 2016). Python packages PyTorch and Scikit-learn were used to test two types of supervised machine learning algorithms: Transfer learning with BERT and Random Forest.

Depending on the algorithms, the labeled dataset was split differently in the training process. For Random Forest, 80% of the dataset was sampled randomly for 5-fold cross validation while the rest was used for testing. For BERT, the dataset was split into three parts: training, validation, and testing sets with a ratio of 60%, 20%, and 20% respectively. Specifically, 60% of the labeled data were used to train the BERT model to detect features and classify posts, 20% which were not included in the training set were used to validate the BERT models (e.g., tuning the parameters to avoid overfitting), and

the rest of the 20% served as testing data to compare the labels generated by the text classifier and the labels given by human coders.

Metrics such as accuracy, precision, recall, F-measure, and Matthew's correlation coefficient were used to show the performance of each model (see Table 8). A comparison among the candidate models revealed that, the model built on BERT and NER features yielded the best performance over all other models. Therefore, this model was chosen to classify the rest of the unlabeled texts from the six forums of the MOOC.

Table 8: Performance Evaluation of the Candidate Models for Text Classification

Models	Features	Accuracy	Precision	Recall	F1	Matthew's correlation coefficient
BERT	Raw text	0.80	0.81	0.75	0.77	0.77
BERT	POS	0.80	0.81	0.78	0.79	0.77
BERT	**NER**	**0.83**	**0.82**	**0.80**	**0.81**	**0.81**
BERT	POS + NER	0.80	0.79	0.78	0.79	0.77
Random Forest	Raw text	0.33	0.57	0.24	0.23	0.30
Random Forest	POS	0.37	0.52	0.27	0.26	0.33
Random Forest	NER	0.39	0.62	0.29	0.30	0.36
Random Forest	POS + NER	0.44	0.65	0.35	0.35	0.41
Random Forest	WE	0.56	0.57	0.46	0.46	0.51
Random Forest	POS + WE	0.51	0.59	0.41	0.41	0.45

Table 8, continued.

Random Forest	NER + WE	0.55	0.57	0.45	0.45	0.49
Random Forest	POS + NER + WE	0.51	0.53	0.40	0.40	0.44

Analyzing the changes of social presence in response to the evolvement of the learner network

As mentioned in the last chapter, the massive and complex nature of MOOCs results in a wide range of participation patterns, with learners randomly dropping in and out in the forums. Furthermore, since each module in the MOOC dealt with different topics/concepts with varying levels of complexity, the dynamics of interaction among learners may change across different modules, which were reflected in the changes of density of the learner network, and may respond to the changes of social presence over time. For instance, at the beginning of the MOOC, learners may engage in more self-disclosure as they introduce themselves to the whole community; as the instructor proceeded to introduce more novel concepts, learners may ask more questions, or express more negative emotions while encountering complicated concepts that are difficult to digest shortly. In contrast, more consensus may be reached among learners when the concepts are relatively easy to comprehend; while for modules explaining procedural knowledge with more hands-on practice, learners may give more advice to each other by commenting on the details of one's solution to a task, as well as share more recourses to show others what external materials or tools are used to accomplish the tasks. To better capture these nuances of participation over time as the MOOC proceeds, this study broke down the network of the entire course by looking at the network of each module. Within

each module, SNA was adopted to examine the network of passive participation and active participation respectively. Passive participation, in this study, refers to the act of reading posts; while active participation is defined as creating posts in the forums. Five different network parameters were used to measure the distinct features of the network of both passive and active participation in each module: network density index, number of nodes, number of edges, number of groups within the network, and group size. Learners' social presence were also analyzed within each module to see how learners present themselves differently as the topics of discussions or learning tasks changed over time. This also provides important contextual information to understand what types of social presence constitute a tightly or sparsely connected learner network.

Examining the relationship between learners' social presence and their network centrality

The learners' centrality in the forums was measured using four SNA parameters: in-degree centrality, closeness centrality and betweenness centrality and Eigen centrality. Specifically, in-degree centrality is determined by the total number of replies one receives from others. High in-degree centrality indicates that others interact frequently with this particular participant in the network. This might imply, for example, that the participant is a popular student in the network or that the nature of his or her posts are in some way interesting or remarkable from the others' point of view. By contrast, closeness centrality measures the average distance from one node to all other nodes in the network (Wasserman & Faust, 1994). In the context of this study, a learner with low closeness centrality means he or she has closer ties with other learners in the forums. Betweenness centrality, on the other hand, shows how often a given participant is found in the shortest path between two other participants in the network, implying how often one learner

serves as a bridge of communication between other two learners (Wasserman & Faust, 1994). Finally, Eigen Centrality measures a node's influence based on the number of links it has to other nodes within the network. Eigen centrality then goes a step further by also taking into account how well connected a node is, and how many links their connections have through the network (Wasserman & Faust, 1994). All centrality calculations were completed in Gephi (Bastian, Heymann, & Jacomy, 2009), an open-source network analysis and visualization software.

To investigate the link between learners' social presence and network centrality, correlation and multiple regression were chosen for statistical analysis, with the 13 social presence indicators as independent variables and the four centrality parameters as dependent variables. Besides the qualitative features of the posts, this study also hypothesizes that learners' posting behaviors may add additional predicting power to learners' network status. In other words, learners' posting behaviors might also affect how central they are in the network. Therefore, three quantitative measures of learners' posting behaviors (frequency of posting, average length of posts, posting day) were included in the regression analysis to see what role learners' posting behaviors played in determining their network status (Figure 4). Specifically, frequency of posting was measured by the total number of posts one had contributed to the forums over the six modules, which is captured by their out-degree in the course. Average length of posts means the average word count of posts. And posting day was determined by the difference of days between the date a post was published on and the date the very first post was created in that forum by a learner (e.g., the very first post in a forum is marked as "0", then the post created three days after will be marked as "3"). The purpose of using mixed methods is to see whether the qualitative nature of the posts, in combination with learners' posting behaviors, predict learners' network status in the learning community.

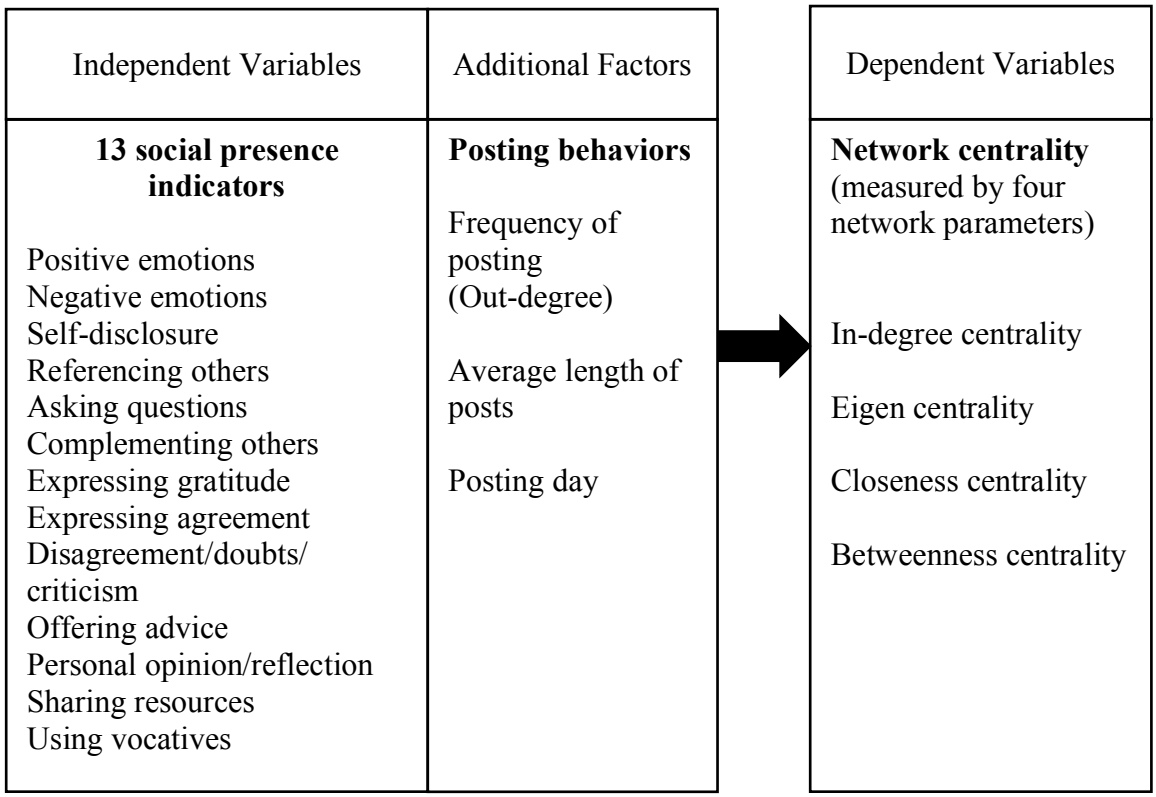


Figure 4: The Independent Variables, Additional Factors and Dependent Variables of Regression Analyses

Examining the correlation between learners' network centrality and their learning outcome

To investigate the correlation between learners' network centrality and their learning outcome, correlation analyses were conducted with learners' network status (measured by the four centrality parameters) as independent variables, and their learning outcome (certificate status, perceived learning, satisfaction) as dependent variables.

Chapter 4: Results

This study adopts mixed methods to explore both the qualitative (learners' social presence) and the quantitative (learners' posting behaviors) nature of learners' forum posts, in order to understand how these features predict learners' centrality in the learner network and their learning outcomes (measured by certificate status, perceived learning, and satisfaction) in a MOOC. The data analysis aims to answer the following research questions:

1. What social presence did learners exhibit in the discussion forums over the six modules of the MOOC?
2. How does the structure of learner network change over the six modules of the MOOC? And how does the learners' social presence differ as the learner network evolves over time?
3. What is the relationship between learners' social presence and their centrality in the learner network? And how do learners' posting behaviors contribute to the prediction of their centrality?
4. How do learners' network centrality correlate with their learning outcomes?

This chapter reports the results of data analysis that answer each of these research questions.

THE DISTRIBUTION OF SOCIAL PRESENCE

Research question 1 requires the examination of social presence learners exhibited over the six modules. Table 9 presents the distribution of the number of each social presence indicator across six modules. In particular, "Number of Comments required" means the minimum number of posts a learner was required to make by commenting on others' assignments. This does not include the learners' own response to the tasks of each

forum. “Number of participants” indicates the number of participants in the forum in each module. Figure 5 gives a more intuitive presentation of the fluctuation of the 13 social presence indicators across the six modules.

As presented in the Table 9 and Figure 5, despite the significant loss of participants in Module 2, the number of sentences for *Expressing agreement* still increased by 79% from 154 in Module 1 to 276 in Module 2, implying that participants reached more consensus in the topic of preparing data in Module 2 than finding and getting data in Module 1. While for Module 3, where students were not required to comment on others’ posts other than posting their own responses to the task, all social presence indicators decreased except for *Referencing others*, which increased from 35 to 55 in Module 3. Another salient finding was that in Module 4, where commenting on others was not mandatory, six social presence indicators still increased, which were *Asking questions, Disagreement/doubts/criticism, Expressing agreement, Negative emotions, Offering advice, Self-disclosure*. The boost of these social presence indicators could possibly relate to the topic—*machine learning in data journalism*, which is popular yet controversial trend in the field of journalism. Students were expected to report their personal experience of using machine learning in their work, the caveats of it and potential solutions, which triggered heated discussions among the forum participants. For instance, *Asking questions* increased 131% from 88 in Module 3 to 204 in Module 4; *Offering advice* increased 46% to 974; *Self-disclosure* increased 40% to 246; and there was also a 26% increase, from 190 to 241, on *Expressing agreement* in Module 4. There was not a radical increase in *Disagreement/doubts/criticism*, but it still rose to 400 despite the loss of participants. The nature of these social presence indicators indicates that students were actively engaging in the discussions, reflecting on their own experience, asking questions, expressing concerns/doubts, debating ideas and offering advice to each

other to address the problems they encountered. Interestingly, Module 4 was the only module where the expression of *Disagreement/doubts/criticism* and *Negative emotions* increased compared to the previous module, even with the loss of participants. This finding implies that a topic triggering conflicting opinions and negative emotions may spark more discussions among learners.

From Module 4 to Module 5, there was an increasing trend of positive social presence indicators such as *Complimenting others* (from 471 to 1503), *Expressing gratitude* (from 129 to 675) and *Positive emotions* (from 454 to 679). A possible explanation of the growth may be, in this module, students were asked to share their data visualization assignments in the forum and were required to make at least four comments to peers. The increase of these positive social presence represents an inclusive and welcoming atmosphere within the forum where students showed respect and appreciation when critiquing each other's work. Another social presence indicator that contributed to the positive atmosphere for learning in this module, was *Using vocatives*, which increased by 48% (from 634 to 940). Referring to each other by name shows a tendency to build social ties with others. Therefore, the surge of *Using vocatives* implies the development of a more friendly and welcoming learning community as the course progressed. Within this favorable environment, students also shared more *Personal Opinion/reflection* (from 108 to 285) than in the previous module. In the meantime, there were more posts of *Referencing others* (from 31 to 58) in Module 5, indicating that learners read and quoted more of their peers' posts than in Module 4. Another finding that is worth noting is the boost of *Sharing resources*, which increased by 46% from 742 to 1085. This is expected since students were asked to share the links of their assignments and other resources they used to complete the assignments. Moving to Module 6, where students progressed to the final step of their incremental data visualization assignments,

there was a 28% increase in *Self-disclosure* (from 233 to 298) and a minor growth in *Positive emotions* (from 679 to 698). Since most students chose to create their data visualization project of personal relevance, the final step of interpreting their projects is expected to involve more *Self-disclosure*. On the other hand, the *Positive emotions* mostly came from students' self-reflection of accomplishing their final projects, and their positive comments on the course and the instructors.

Table 9: The Distribution of Social Presence Over the Six Modules

Topic of each module	M1 Finding and getting data	M2 Preparing data	M3 Finding stories in data	M4 Machine learning in journalism	M5 Visualizing data	M6 Data-driven story - telling
Number of Comments required	2	3	0	0	4	4
Sharing links required	Yes	Yes	Yes	No	Yes	Yes
Number of participants	1789	908	311	282	250	227

Table 9, continued.

Asking questions	758 (4.9%)	268 (2.9%)	88 (1.6%)	**204 (4.0%)**	190 (2.7%)	180 (3.1%)
Complimenting others	1867 (12.0%)	1147 (12.3%)	891 (16.4%)	471 (9.3%)	**1305 (18.5%)**	1114 (18.9%)
Disagreement/doubts/criticism	250 (1.6%)	478 (5.1%)	194 (3.6%)	**200 (4.0%)**	175 (2.5%)	81 (1.4%)
Expressing agreement	154 (1.0%)	**276 (3.0%)**	190 (3.5%)	**241 (4.8%)**	182 (2.6%)	178 (3.0%)
Expressing gratitude	882 (5.7%)	531 (5.7%)	215 (4.0%)	129 (2.6%)	**675 (9.6%)**	669 (11.3%)
Negative emotions	1181 (7.6%)	1173 (12.6%)	533 (9.8%)	**607 (12.0%)**	607 (8.6%)	452 (7.7%)
Offering advice	1662 (10.7%)	1187 (12.7%)	666 (12.2%)	**974 (19.3%)**	637 (9.0%)	377 (6.4%)
Personal opinion/reflection	1012 (6.5%)	671 (7.2%)	212 (3.9%)	108 (2.1%)	**285 (4.0%)**	163 (2.8%)
Positive emotions	1780 (11.4%)	647 (6.9%)	639 (11.8%)	454 (9.0%)	**679 (9.6%)**	**698 (11.8%)**
Referencing others	70 (0.5%)	35 (0.4%)	55 (1.0%)	31 (0.6%)	**58 (0.8%)**	39 (0.7%)

Table 9, continued.

Self-disclosure	908	325	175	**246**	233	**298**
	(5.8%)	(3.5%)	(3.2%)	**(4.9%)**	(3.3%)	**(5.1%)**
Sharing resources	2710	1330	880	742	**1085**	909
	(17.4%)	(14.3%)	(16.2%)	(14.7%)	**(15.4%)**	(15.4%)
Using vocatives	2314	1260	699	634	**940**	737
	(14.9%)	(13.5%)	(12.9%)	(12.6%)	**(13.3%)**	(12.5%)

Note. "Number of Comments required" means the minimum number of posts a learner was required to make by commenting on others' assignments. This does not include the learners' own response to the tasks of each forum. The numbers in bold indicate an increase in that social presence indicator compared to the previous module. The percentages in the brackets indicate the percentage of a particular social presence indicator in all posts of the module.

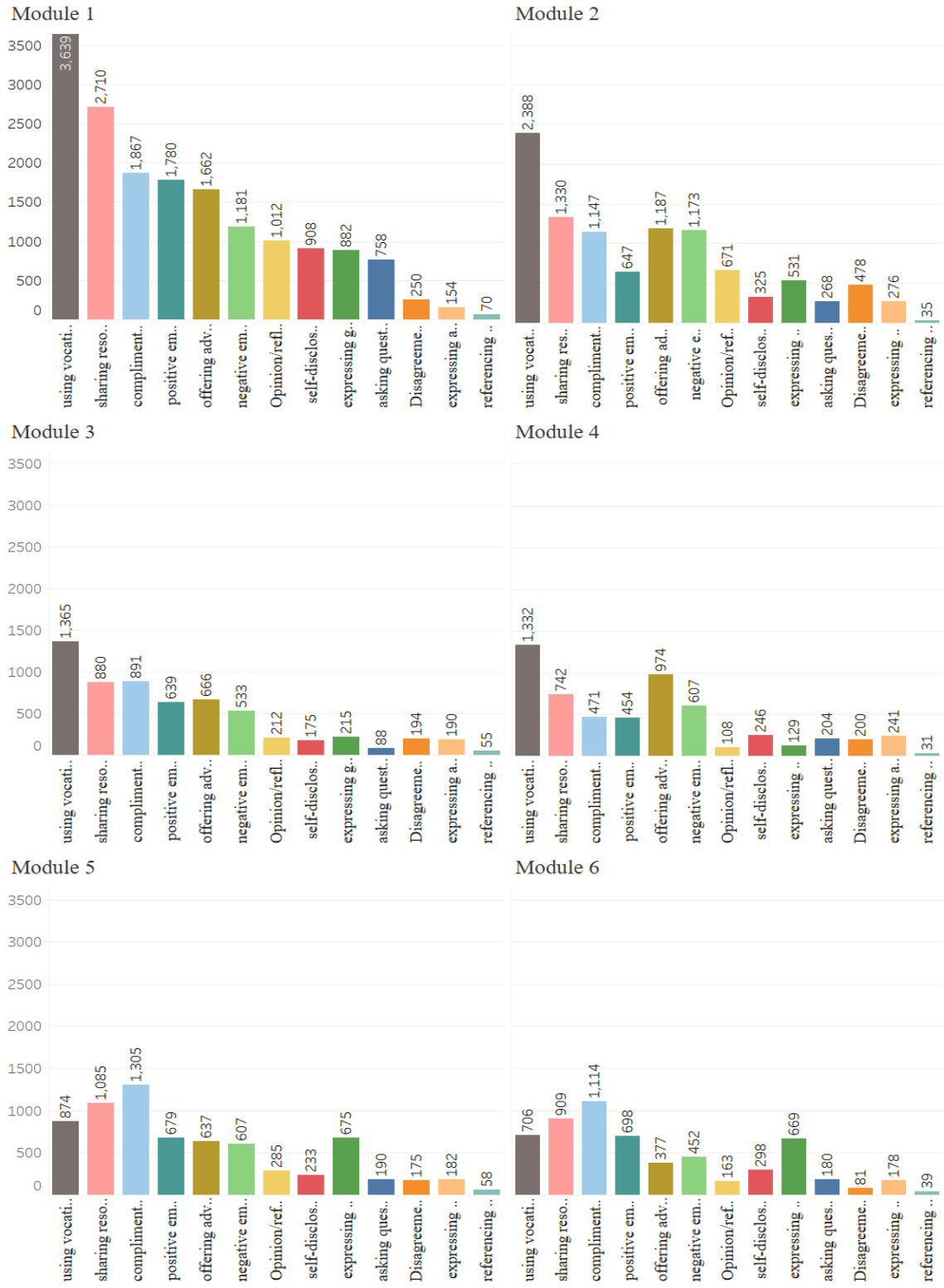

Module 1
3500
3000
2500
2000
1500
1000
500
0
3,639
2,710
1,867
1,780
1,662
1,181
1,012
908
882
758
250
154
70
using vocati..
sharing reso..
compliment..
positive em..
offering adv..
negative em..
Opinion/refl..
self-disclos..
expressing g..
asking quest..
Disagreeme..
expressing a..
referencing ..
Module 2
2,388
1,330
1,147
647
1,187
1,173
671
325
531
268
478
276
35
using vocat..
sharing res..
compliment..
positive em..
offering ad..
negative e..
Opinion/ref..
self-disclos..
expressing ..
asking ques..
Disagreeme..
expressing ..
referencing ..
Module 3
1,365
880
891
639
666
533
212
175
215
88
194
190
55
Module 4
1,332
742
471
454
974
607
108
246
129
204
200
241
31
Module 5
874
1,085
1,305
679
637
607
285
233
675
190
175
182
58
Module 6
706
909
1,114
698
377
452
163
298
669
180
81
178
39

Figure 5: The Distribution of Social Presence Over the Six Modules

LEARNERS' PARTICIPATION PATTERNS IN THE DISCUSSION FORUMS IN RELATION TO THEIR SOCIAL PRESENCE

Research question 2 guided the analysis of learners' participation pattern, and the changes of social presence in relation to the evolvement of the learner network. Results are presented below.

Passive participation

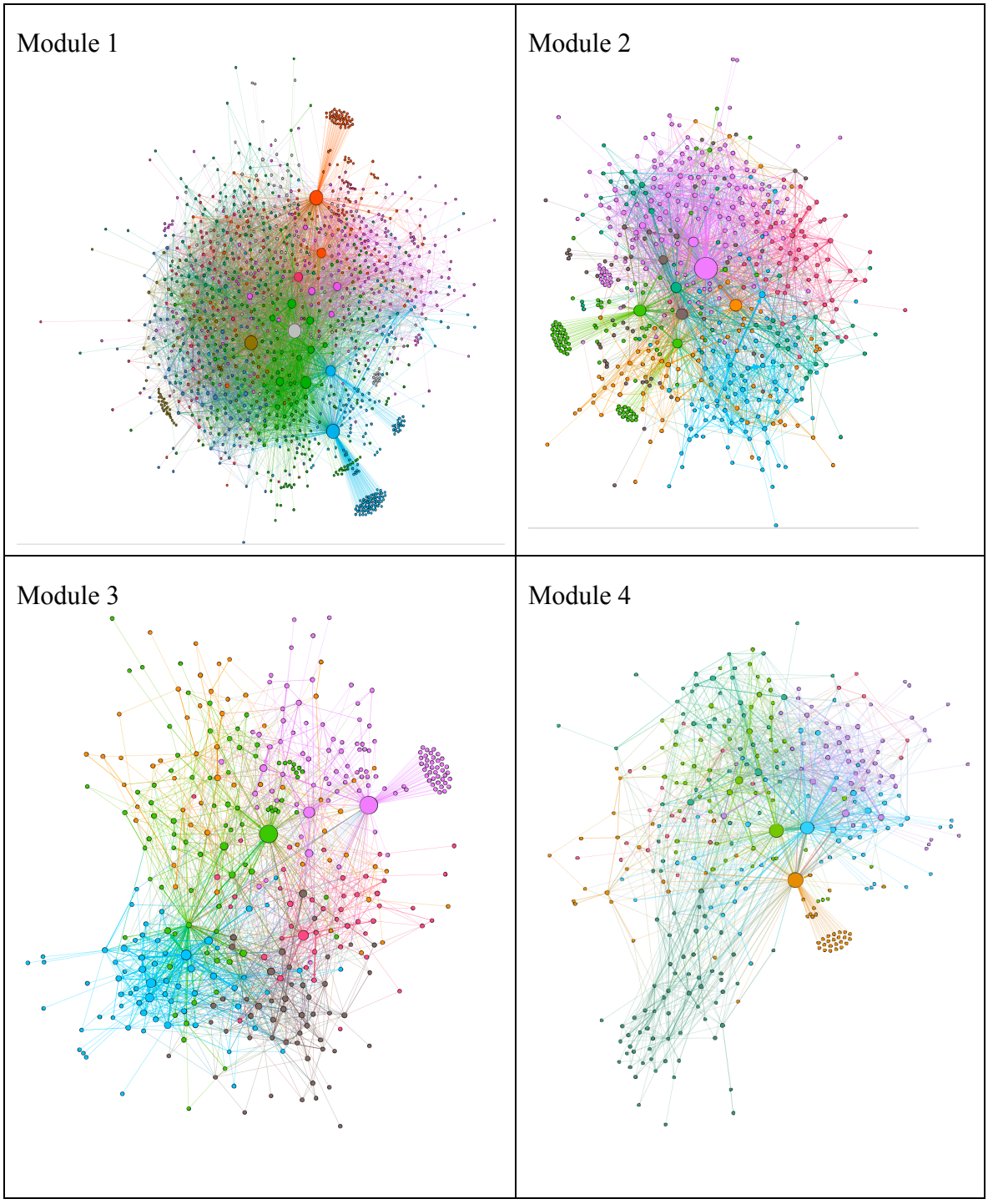
The learner network evolved over the six modules as learners dropped in and dropped out. Table 10 and Figure 6 shows the network patterns of learners' passive participation (viewing without posting) for each module. A comparison of the six networks in Table 10 revealed that the number of learners in the network gradually declined as the MOOC progressed over time. From Module 1 to Module 2, there was a drastic decrease of participation in the discussion forum with more than half of the learners (n = 710) opted out. While from Module 2 to Module 3, around 26% of the learners (n = 148) left the forum. As the course moved on to Module 4, the declining trend starts to slow down. There was a loss of 17% of the learners (n = 68) from Module 3 to Module 4, while the participation rate stayed almost the same from Module 4 to Module 5, and a loss of 64 learners (18%) from Module 5 to Module 6. In terms of network density, as the learner network shrunk in every module (due to the loss of learners), the network became gradually tighter, as indicated in the network density index in Table 10, although this can be partly explained by the fact that there were less participants to respond to/interact with in the discussions, it could also imply that the

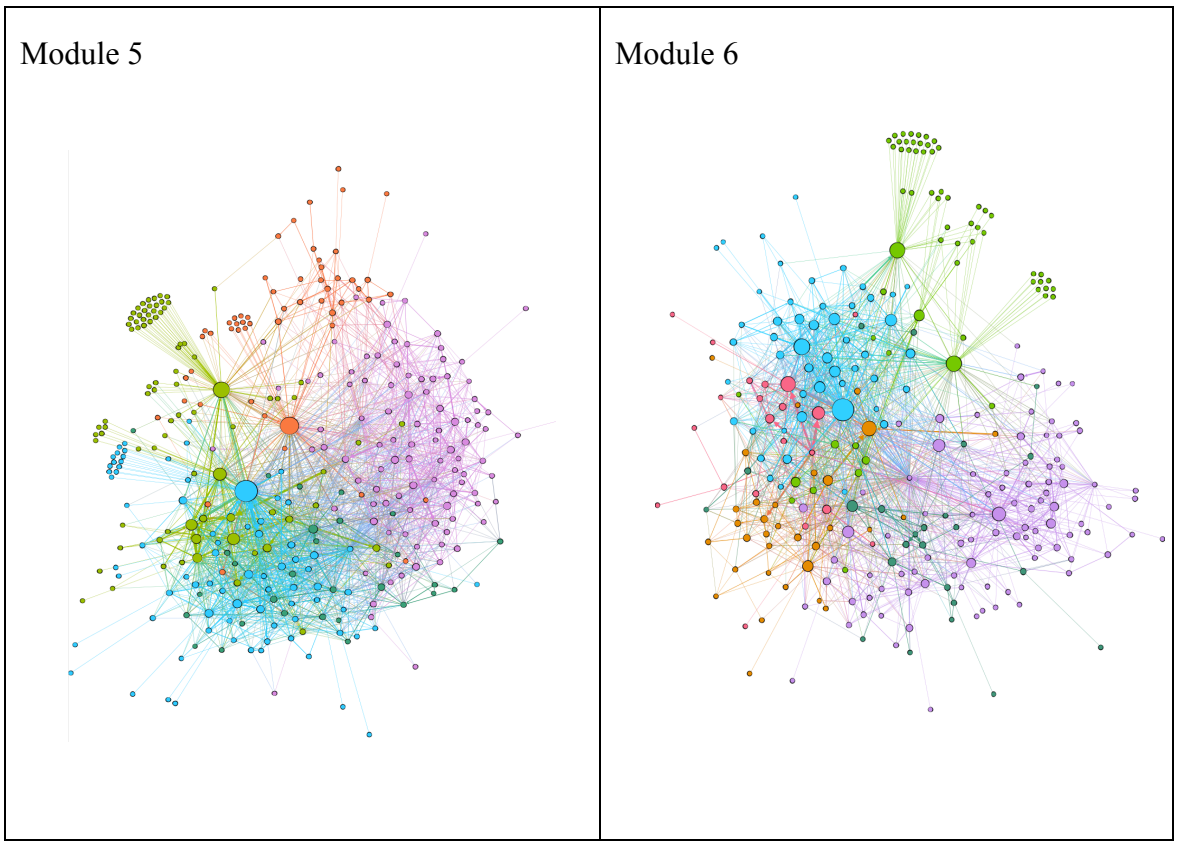
general level of connections among the learners grew higher by module, By the time of Module 6, the learner network became the most densely interconnected compared to that of the previous five modules (Figure 6). When looking at the frequency of learners' latent interaction with one another through reading posts, which is captured by the number of edges in the network, it appeared that the number peaked at Module 1 with a total of 7518 interactions, with the highest number of participants in the first module. After more than half of the learners dropped out from the forum in Module 2, the number of edges also dropped by 62% to 2871. The latent interactions continued to drop in Module 3, although not as dramatically as in the previous module. When the course progressed to Module 4, there was a slight uptick in edges (6%) compared to Module 3. It continued to grow in Module 5 to 2082 and finally dropped to 1626 in Module 6, which was the lowest level of latent interactions among the six modules.

When learners engage in the conversations in discussion forums, they often form sub-groups or communities - some are bigger while others are smaller with less participants. The calculation of modularity yielded the number of groups in the learner network in each module. There was a clear trend of decrease in the number of groups as the course progressed (from 10 groups in Module 1 to only 5 groups in Module 6). This is partly due to the decrease of participants over the six modules. In terms of the size of groups, the number of participants in each group became smaller over time. In other words, the size of groups shrunk as more learners left the forum. The only exception was Module 4, in which the maximum size of the group grew more than twice larger than that in Module 3, indicating that there were some larger groups with tightly connected individuals than in the previous module. In other words, there may be some very popular posts/threads that attracted a lot of learners to read.

Table 10: The Learner Network of Passive Participation Over the Six Modules

	Network density	Number of nodes	Number of edges	Number of groups in the network	Size of groups
Module 1	0.005	1269	7518	10	30-320
Module 2	0.009	559	2871	7	50-160
Module 3	0.011	411	1782	7	35-130
Module 4	0.016	343	1894	6	10-280
Module 5	0.018	345	2082	5	30-100
Module 6	0.021	281	1626	5	20-90


Module 1
Module 2
Module 3
Module 4



Note. The same color of nodes in the social diagrams indicates that they are in the same group in the learner network. The larger nodes in different networks may represent different learners.

Figure 6: The Network Diagrams of Learners' Passive Participation Over the Six Modules

Active participation

Table 11 and Figure 7 show the network patterns of learners' active participation (posting and replying others) for each module. Similar to passive participation, Table 11 revealed that the number of participants in the network gradually declined as the MOOC progressed over time (from 1789 participants in Module 1 to 227 participants in Module 6). From Module 1 to Module 2, there was a drastic decrease of active participation in the forum with 49% of the learners (n = 881) opted out. While from Module 2 to Module 3,

there was an even more precipitous drop of 66% participants (n = 597). As the course moved on to Module 4, the declining trend started to slow down. There was a loss of 9% of the learners (n = 29) from Module 3 to Module 4, a loss of 11% of the learners (n = 32) from Module 4 to Module 5, and another loss of 11% the learners (n = 23) from Module 5 to Module 6. The network density closely resembles the patterns found in passive participation. As the learner network shrunk in every module (due to the loss of participants), the network became gradually tighter, as indicated in the network density measures in Table 11, implying that the general level of connections among the learners grew higher by module. By the time of Module 6, the learner network became the most densely interconnected compared to that of the previous five modules. When looking at the frequency of learners' explicit interactions through composing and replying posts, which is captured by the number of edges in the network, it appeared that the most active interactions also occurred in Module 1 (with a total of 2427 interactions), with the highest number of participants in the first module. After almost half of the learners dropped out from the forum in Module 2, the number of edges also dropped in half to 1268. With the most dramatic decrease in participants in Module 3, the interactions continued to drop in half (53%) to 597 in Module 3. When the course progressed to Module 4, the drop of edges slowed down, losing only 6% of the interactions from Module 3. There was an uptick of interaction with an increase of 297 edges in Module 5, then it decreased again to 712 edges in Module 6. Unlike the passive participation network, the lowest interaction occurred in Module 4 in terms of active participation.

It is interesting to note that, similar to the passive participation network, as the overall size of the network became smaller, the number of groups in the network also decreased, from 144 groups in Module 1 to 33 groups in Module 6. As the network became tighter, the size of the groups also shrunk. For example, the maximum size of

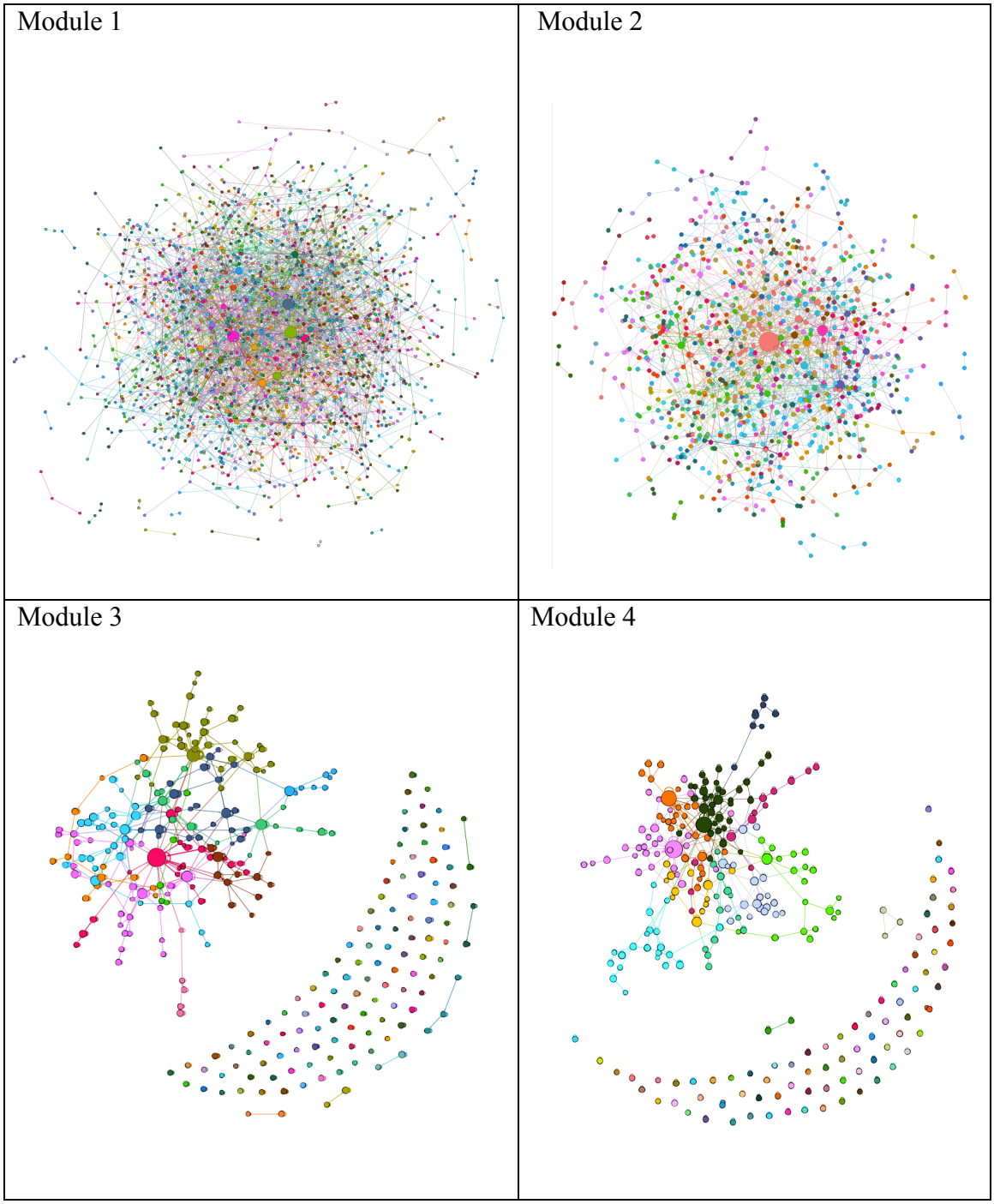
group in Module 1 was 150, while this number dropped by more than half in Module 2, with the largest group having only 72 members. Interestingly, in Module 3, the number of groups doubled from that in Module 2 (n = 67) to 125, but with group size significantly smaller than that in the previous module. This is possibly due to the fact that there was no requirement of commenting on others' posts in Module 3. Figure 7 reflects this change in the network graph, with a large number of isolated individuals not interacting with any peers. This pattern continued in Module 4, with a large number of isolated "groups" (n = 91) but small group size (group size =1, meaning these "groups" actually consisted of only one individual). However, when the course progressed to Module 5, there was a precipitous drop in the number of groups. While the number of participants remained similar to the previous module, the network density in Module 5 doubled with fewer groups but tighter connections among participants. The pattern of Module 6 is very close to that of Module 5 in terms of the number of participants, network density and the number of groups in the network, but with smaller group size. In general, the network patterns in active participation varied more than that in the passive participation. Especially in Module 3 and 4, where commenting on others' posts was not mandated.

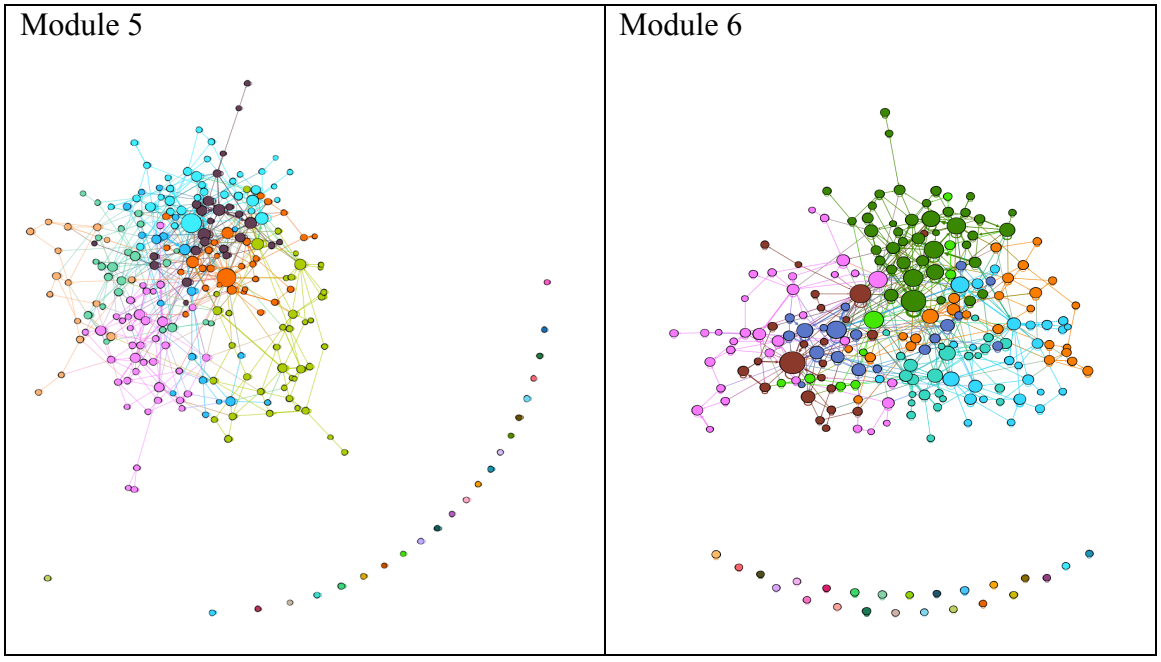
In Module 1 and Module 2, which required finding external resources, sharing links and commenting on peers' posts in the forums, the distribution of social presence reflects that the nature of students' posts was aligned with the requirements of the forums, with high occurrences of *Sharing resources, Using vocatives, Complimenting others* and *Offering advice*. The high frequencies of *Using vocatives, Complimenting others* and *Offering advice* indicate that the nature of the posts was interactive and conversational, which is captured by the high number of edges in the network. However, due to the large number of participants in the first two modules, the networks were still relatively sparse. By contrast, in Module 3 and 4, where sharing external information and

commenting on others were both optional, *Sharing resources, Using vocatives, Complimenting others* and *Offering advice* still dominated the discussions. This suggests that even though conversations with others was not mandatory, learners were still proactive in sharing useful links to external resources and initiating discussion with peers in Module 3 and 4. The increased network density in these two modules reflects this high level of interaction. However, as demonstrated in Figure 7, there were still a great deal of learners who merely posted their own thoughts without interacting with any peers. This was captured by the high number of groups and smaller group size in the network in these two modules (most "groups" only consisted of one participant). When the course moved to Module 5 and 6, when learners made more progress in their incremental data visualization projects and more comments were required to critique each other's work, *Sharing resources, Using vocatives, Complimenting others* and *Expressing gratitude* became the most prominent social presence indicators in the forums. The network became denser as the community became more stable with a group of persistent learners. Even with a loss of participants from previous modules, the conversational nature of the learners' posts and the high number of edges signaled a tight-knit community with devoted participants.

Table 11: The Learner Network of Active Participation Over the Six Modules

	Network density	Number of nodes	Number of edges	Number of groups in the network	Size of groups
Module 1	0.001	1789	2427	144	1-150
Module 2	0.002	908	1268	67	1-72
Module 3	0.006	311	597	125	1-28
Module 4	0.007	282	561	91	1-35
Module 5	0.014	250	858	32	1-46
Module 6	0.014	227	712	33	1-36


Module 1
Module 2
Module 3
Module 4



Note. The same color of nodes in the social diagrams indicates that they are in the same group in the learner network; the isolated nodes represent those who didn't comment on any peers' posts and never received any comments from peers.

Figure 7: The Network Diagrams of Learners' Active Participation Over the Six Modules

THE RELATIONSHIP BETWEEN LEARNERS' SOCIAL PRESENCE AND NETWORK CENTRALITY

Research question 3 guided the investigation of the relationship between learners' social presence and network centrality. Specifically, this study attempts to explore the correlation between learners' social presence and network centrality, as well as how learners' social presence predicts their network centrality, and what role learners' posting behaviors (measured by the frequency of posting, average length of posts, posting day) play in the prediction. The results are presented as follows.

The correlations between social presence and learners' centrality

Since the data was not normally distributed, Spearman's correlation analyses were selected to detect the correlation between social presence and learners' centrality in the network measured by in-degree, Eigen centrality, closeness centrality and betweenness centrality. The results are presented in Table 12.

In-degree value represents the number of replies a learner receives in the discussion forums. Results showed that, among the 13 social presence indicators, *Expressing agreement* (r = .521) had the strongest correlation with in-degree, meaning that a learner is more likely to receive more replies when his or her posts have higher percentage of *Expressing agreement.* Moderate correlations (.3 < r < .5, Cohen, 1988) were found between in-degree and four social presence indicators, namely *Negative emotions* (r = .325)*, Expressing gratitude* (r = .396)*, Referencing others* (r = .400)*,* and *Disagreement/doubts/criticism* (r = .445). This implies that learners with relatively higher response rate devoted a higher percentage of their posts to express negative emotions, disagreement, gratitude, and address other's opinions in their own posts. Weak correlations were detected between in-degree and five social presence indicators: *Personal opinion/reflection* (r = .136)*, Self-disclosure* (r = .145), *Offer advice* (r = .154)*, Complimenting others* (r = .240)*,* and *Asking questions* (r = .268). This implies that although these types of social presence may trigger more responses from others, the likelihood is relatively low.

Eigen centrality is conceptually similar to in-degree, but with an emphasis on quality of one's connection in the network. Therefore, the pattern of correlations between Eigen centrality and social presence is similar to that between in-degree and social presence. One difference was that no strong correlation was found (r > .5, Cohen, 1988). While moderate correlations occurred between Eigen centrality with four social presence

indicators: *Referencing others* (r = .328)*, Disagreement/doubts/criticism* (r = .360), *Expressing gratitude* (r = .374)*,* and *Expressing agreement* (r = .451). Weak correlations were detected between in-degree and six social presence indicators: *Personal opinion/reflection* (r = .104)*, Self-disclosure* (r = .106)*, Offer advice* (r = .141)*, Complimenting others* (r = .253)*, Asking questions* (r = .272), and *Negative emotions* (r = .292).

Closeness centrality calculates the shortest paths between one node and all other nodes in a network, in order to identify the individuals who are best placed to influence the entire network. Similar to the patterns described above, *Expressing agreement* (r = .537), *Expressing gratitude* (r = .472) and *Disagreement/doubts/criticism* (r = .331) still stood out to be the most important social presence indicators that had stronger correlations with learners' closeness centrality. This indicates that taking stance (agreement and disagreement) and showing gratitude to others are more likely to push a learner towards a favorable position to influence the whole network. *Complimenting others* (r = .314) was found to have a slightly stronger correlation with one's closeness centrality than with in-degree and Eigen centrality. *Negative emotions* (r = .177)*, Asking questions* (r = .254) *and Referencing others* (r = .282), on the other hand, appeared to have weak correlations with closeness centrality.

Betweenness centrality calculates how many times a node falls on the shortest paths between other nodes. It helps to identify the individuals who influence the information flow within a network. Again, *Expressing agreement* (r = .485), *Expressing gratitude* (r = .404) and *Disagreement/doubts/criticism* (r = .337) still stood out to have the strongest link with learners' betweenness centrality.

Weak correlations were detected between betweenness centrality and seven social presence indicators: *Referencing others* (r = .289), *Negative emotions* (r = .281),

Asking questions (r = .274), *Complimenting others* (r = .234), *Self-disclosure* (r = .111), *Offering advice* (r = .108) and *Using vocatives* (r = .081).

Interestingly, *Sharing resources* was the only social presence indicator that had significant negative correlations with all four centrality measures, though these negative associations were relatively weak (r < .3). This implies that when a learner devotes a higher percentage of his or her posts to sharing URLs, it is detrimental to his centrality in the network. Given the context of this MOOC, students were frequently required to give the links of examples/one's own assignments in the forum. But if one's social presence in the forum is limited to providing URLs without much effort to initiate other types of interaction with others, it is reasonable that he or she will become a peripheral participant in the network.

Table 12: The Correlations Between Learners' Social Presence and Network Centrality

	In-degree	Eigen	Closeness	Betweenness
Asking questions	.268**	.272**	.254**	.274**
Complimenting others	.240**	.253**	.314**	.234**
Disagreement/doubts/criticism	.445**	.360**	.331**	.337**
Expressing agreement	.521**	.451**	.537**	.485**
Expressing gratitude	.396**	.374**	.472**	.404**
Negative emotions	.325**	.292**	.177**	.281**
Offering advice	.154**	.141**	0.07	.108**
Personal opinion/reflection	.136**	.104**	-0.038	0.027
Positive emotions	0.012	-0.021	-0.022	0.025

Table 12, continued.

Referencing others	.400**	.328**	.282**	.289**
Self-disclosure	.145**	.106**	0.059	.111**
Sharing resources	-.215**	-.237**	-.181**	-.250**
Using vocatives	0.045	0.048	0.073	.081*

Note. **. $p < .01$; *. $p < .05$ (corrected by the sequential Bonferroni method).

Predicting network centrality from social presence and posting behaviors

Prior to conducing hierarchical multiple regression analysis, the relevant assumptions of this statistical analysis were tested. Firstly, a sample size of 721 (the total number of those who posted at least once) was deemed adequate given 16 independent variables (13 social presence indicators and 3 posting behavior factors) to be included in the analysis (Tabachnick & Fidell, 2001). The examination of collinearity revealed that *Sharing resources* as a predictor has a tolerance of 0, meaning that the variance in *Sharing resources* is already contained in, or is redundant with other predictors. Therefore, it was excluded from the regression model to meet the assumption of multicollinearity (Coakes, 2005; Hair, 1998).

A four-step hierarchical multiple regression was conducted with each of the four SNA centrality measures as the dependent variable. At step one, the 12 social presence indicators (after excluding *Sharing resources*) were entered into the regression model as independent variables. Out-degree was entered into the model at step two, while average length of post at step three and posting day at step four. The regression statistics are reported in Table 13, 14, 15 and 16, with in-degree, Eigen centrality, closeness centrality and betweenness centrality as dependent variables respectively.

In the regression models that predict in-degree (Table 13), model 1 showed that the seven social presence indicators (*Complimenting others, Disagreement/doubts/criticism, Expressing agreement, Expressing gratitude, Negative emotions, Offering advice, Referencing others*) contributed significantly to the regression model, F (12,708) = 7.5, $p < .05$, and accounted for 11.3% of the variation in in-degree. Introducing out-degree in the prediction in step two explained an additional 67.3% of the variations in in-degree. And this change in R^2 was significant. In step three, adding the *average length of post* as a predictor to the regression model only explained an additional 0.1% of the variations in in-degree and this change in R^2 was not significant. Finally, the addition of *posting day* as a predictor to the regression model explained an additional 0.7% of the variations in in-degree and this change in R^2 was also significant. In model 2 and model 3, *Expressing agreement* and *out-degree* were the only two predictors that contribute significantly to the model. Whereas in model 4, besides *Expressing agreement* and *out-degree*, *posting day* also appeared to be a significant predictor to the model. But *out-degree* remained the most important predictor of in-degree. When all 15 independent variables were included in model 4, together they accounted for 79.4% of the variances in in-degree.

In the regression models that predict Eigen centrality (Table 14), model 1 showed that the nine social presence indicators (*Asking questions, Complimenting others, Expressing agreement, Disagreement/doubts/criticism, Expressing gratitude, Negative emotions, Offering advice, Referencing others, Using vocatives*) contributed significantly to the regression model, F (12,708) = 6.625, $p < .05$, which in total accounted for 9.3% of the variation in Eigen centrality. Introducing out-degree in the prediction in step two explained an additional 51.4% of the variations in Eigen centrality. And this change in R^2 was significant. Compared to step one, only *Asking questions* and *out-degree* remained

the significant predictor of Eigen centrality in step two. In step three, including the *average length of post* as a predictor to the regression model added no predicting power to the variations in Eigen centrality. Finally, the addition of *posting day* as a predictor to the regression model explained an additional 3.4% of the variations in Eigen centrality and this change in R^2 was significant. But among all 14 predictors in model 3, *out-degree* was the only predictor that contributed significantly to the model. Whereas in model 4, both *out-degree* and *posting day* were significant predictors to the model. When all 15 independent variables were included in model 4, together they accounted for 64.1% of the variances in Eigen centrality.

In the regression models that predict closeness centrality (Table 15), model 1 showed that the five social presence indicators (*Asking questions, Complimenting others, Expressing agreement, Expressing gratitude, Using vocatives*) contributed significantly to the regression model, F (12,708) = 7.687, $p < .05$) and accounted for 11.5% of the variations in closeness centrality. Introducing out-degree in the prediction in step two explained an additional 6.6% of the variations in closeness centrality. And this change in R^2 was significant. Besides the five social presence indicators in step one, *out-degree* became another significant predictor of closeness centrality in model 2. In step three, including the *average length of post* as a predictor to the regression model explained an additional 0.2% of the variations in closeness centrality, but this change in R^2 was not significant. Finally, the addition of *posting day* as a predictor to the regression model made no difference to the model. When all 15 independent variables were included in model 4, the aforementioned five social presence indicators and *out-degree* were significant predictors, together they accounted for 18.3% of the variances in closeness centrality.

In the regression models that predict betweenness centrality (Table 16), model 1 showed that only four social presence indicators (*Complimenting others, Expressing agreement, Negative emotions, Offering advice*) contributed significantly to the regression model, F (12,708) = 2.28, $p < .05$, and in total accounted for 3.8% of the variations in betweenness centrality. Introducing out-degree in the prediction in step two explained an additional 54.5% of the variations in betweenness centrality. And this change in R^2 was significant. In model 2, only *Complimenting others, Expressing agreement,* and *out-degree* were significant predictors of betweenness centrality. In step three, including the *average length of post* as a predictor made no difference to the model. Finally, the addition of *posting day* as a predictor to the regression model explained an additional 0.2% of the variations in betweenness centrality, and this change in R^2 was significant. When all 15 independent variables were included in model 4, *Complimenting others, Expressing agreement, out-degree* and *posting day* were significant predictors, together they accounted for 58.5% of the variances in betweenness centrality.

Table 13: Hierarchical Regression Analysis for Predicting In-degree

	Model 1	Model 2	Model 3	Model 4
	β	β	β	β
Step 1 Social presence				
Asking questions	.062	.031	.019	.014
Complimenting others	.183*	-.011	-.022	-.021
Disagreement/doubts/criticism	.116*	.026	.014	.019

Table 13, continued.

Expressing agreement	.167*	.051*	.056*	.055*
Expressing gratitude	.193*	.014	.006	.006
Negative emotions	.172*	.032	.016	.008
Offering advice	.121*	.021	.003	.004
Personal opinion/reflection	.031	.039	.023	.021
Positive emotions	.026	.028	.016	.019
Referencing others	.091*	.012	.008	.009
Self-disclosure	.005	.026	.014	.013
Using vocatives	.065	.023	.008	.004
Step 2				
Out-degree		.901*	.902*	.898*
Step 3				
Average length of post			.037	.033
Step 4				
Posting day				-.084*
R^2	.113	.786	.787	.794
R^2 change		.673*	.001	.007*

Note. *$p < .05$; β: standardized coefficient; n = 721

Table 14: Hierarchical Regression Analysis for Predicting Eigen Centrality

	Model 1	Model 2	Model 3	Model 4
	β	β	β	β
Step 1 Social presence				
Asking questions	.082*	.056*	.048	.037
Complimenting others	.162*	-.007	-.013	-.012
Disagreement/doubts/criticism	.086*	.007	.000	.012
Expressing agreement	.153*	-.038	-.041	-.037
Expressing gratitude	.197*	.040	.036	.036
Negative emotions	.158*	.036	.026	.010
Offering advice	.101*	.014	.004	.006
Personal opinion/reflection	.034	.041	.032	.028
Positive emotions	.019	.021	.013	.020
Referencing others	.077*	.008	.005	.007
Self-disclosure	.008	.027	.019	.017
Using vocatives	.082*	.046	.037	.029
Step 2				
Out-degree		.787*	.788*	.779*
Step 3				
Average length of post			.022	.012
Step 4				

Table 14, continued.

Posting day				-.187*
R^2	.093	.607	.607	.641
R^2 change		.514*	.000	.034*

Note. *$p < .05$; β: standardized coefficient; n = 721

Table 15: Hierarchical Regression Analysis for Predicting Closeness Centrality

	Model 1	Model 2	Model 3	Model 4
	β	β	β	β
Step 1 Social presence				
Asking questions	.112*	.103*	.121*	.120*
Complimenting others	.185*	.125*	.141*	.141*
Disagreement/doubts/criticism	.021	-.007	.012	.013
Expressing agreement	.165*	.097*	.104*	.104*
Expressing gratitude	.200*	.143*	.155*	.155*
Negative emotions	.024	-.020	.005	.004
Offering advice	.053	.021	.049	.049
Personal opinion/reflection	-.032	-.029	-.005	-.005
Positive emotions	.028	.029	.048	.048
Referencing others	-.006	-.031	-.024	-.024
Self-disclosure	-.004	.003	.022	.022
Using vocatives	.144*	.131*	.155*	.155*

Table 15, continued.

Step 2				
Out-degree		.282*	.280*	.280*
Step 3				
Average length of post			-.056	-.056
Step 4				
Posting day				-.008
R^2	.115	.181	.183	.183
R^2 change		.066*	.002	.000

Note. *$p < .05$; β: standardized coefficient; n = 721

Table 16: Hierarchical Regression Analysis for Predicting Betweenness Centrality

	Model 1	Model 2	Model 3	Model 4
	β	β	β	β
Step 1 Social presence				
Asking questions	.049	.023	.018	.016
Complimenting others	.060	-.115*	-.120*	-.118*
Disagreement/doubts/criticism	.055	-.025	-.030	-.026
Expressing agreement	.095*	-.104*	-.106*	-.105*
Expressing gratitude	.144*	-.017	-.020	-.019
Negative emotions	.087*	-.037	-.044	-.047
Offering advice	.084*	-.007	-.014	-.013
Personal opinion/reflection	.011	.017	.011	.010

Table 16, continued.

Positive emotions	.008	.010	.005	.007
Referencing others	.038	-.032	-.034	-.034
Self-disclosure	-.001	.019	.014	.014
Using vocatives	.034	-.005	-.011	-.013
Step 2				
Out-degree		.811*	.811*	.809*
Step 3				
Average length of post			.016	.013
Step 4				
Posting day				-.048*
R^2	.038	.583	.583	.585
R^2 change		.545*	.000	.002*

Note. *$p < .05$; β: standardized coefficient; n = 721

The correlation between learners' centrality and learning outcomes

Research question 4 investigates the correlations between learners' network centrality and their learning outcomes, which were measured by their certificate status, perceived learning and satisfaction. The results are reported as follows.

Table 17 presents the correlation analyses results between the four centrality measures and students' learning outcomes. Specifically, certificate status was found strongly correlated with in-degree (r = .563), closeness centrality (r = .560), betweenness centrality (r = .521), and moderately correlated with Eigen centrality (r = .458). This suggests that learners who received more replies in the forums, had shorter distance with

others, served more frequently as "bridges" between others, and had more high-quality connections were more likely to receive the course certificate in the end. By contrast, learners' centrality appeared to have no association with their perceived learning in any way, as indicated by the correlation results. There was no significant correlation between perceived learning and centrality, or between the four subcategories under perceived learning (namely general learning, cognitive learning, affective learning and behavioral learning) and the four centrality measures. This suggests that learners' centrality in the network is not linked to their perceived learning in any aspect. Finally, when it comes to satisfaction, it is found that only Eigen centrality was significantly correlated with learners' satisfaction, which implies that if a learner has more influential ties in the forums, he or she is more likely to feel satisfied with the learning experience in the MOOC.

Table 17: Correlation Results Between Centrality Measures and Learning Outcome

	In-degree	Eigen Centrality	Closeness Centrality	Betweenness Centrality
Certificate Status	.563**	.458**	.560**	.521**
Perceived Learning	.069	-.007	-.108	.016
General Learning	.065	.050	-.114	-.051
Cognitive Learning	.090	.061	-.029	-.006
Affective Learning	-.018	-.097	-.121	-.071
Behavioral Learning	.048	-.034	-.136	-.142
Satisfaction	.218	.248*	.125	.125

Note. **. $p < .01$; *. $p < .05$ (corrected by the sequential Bonferroni method); n = 71

Chapter 5: Discussion

SUMMARY OF RESEARCH FINDINGS

This study seeks to investigate learners' engagement in MOOC discussion forums from the perspective of social presence. This took place in the context of a MOOC that aims to provide professional development for journalist professionals. The purpose of this study is to understand: first, the distribution of social presence learners exhibited in the discussion forums over the six modules of the MOOC; second, the changes in the learner networks over the six modules in relation to the patterns of social presence; third, how learners' social presence predict their centrality in the learner network, and how their posting behaviors contribute to the prediction; lastly, the correlation between learners' network centrality and their learning outcomes.

To answer the four research questions, this study adopts a mixed-method approach to examine the data from both qualitative and quantitative aspects. To qualitatively analyze the posts, a machine learning enabled text classification model, which was built and validated based on a previous MOOC of the same topic, was adopted to automatically analyze the large-scale text data in the discussion forums. Regression analyses were used in order to understand how learners' social presence and posting behaviors predict their network centrality. Furthermore, correlation analyses were conducted to understand the association between learners' network centrality and their learning outcomes, which were measured by their certificate status, perceived learning and satisfaction. The purpose of using mixed methods is to see whether the qualitative nature of the posts, in combination with learners' posting behaviors, predict learners' position and influence in the learning community and their ultimate learning outcomes.

In the following section, the results are further discussed in the context of the literature. The directions of future research are suggested along with the limitations of this study.

THE DISTRIBUTION OF SOCIAL PRESENCE OVER THE SIX MODULES

To combat the challenge of analyzing large scale of text data, this study adopted a text classifier based on BERT, a latest and revolutionary model in NLP, to analyze forum posts in a MOOC in terms of social presence. The model leverages BERT's exceptional capacity to achieve higher accuracy due to its sensibility to detect bidirectional contextual information of words and its powerful pre-trained data, which is a significant progress compared to traditional methods that isolated words from their contexts (Almatrafi et al., 2018; Wise et al., 2016; Wise, et al., 2017). By comparing the performance of BERT models when adding different linguistic features, this study concluded that BERT with NER yielded the best results. The model developed in this study achieved satisfying performance in a challenging text classification task - classifying text into ten categories, whereas models developed in previous studies were mostly trained to classify text in relatively less categories (typically less than five) (Hu et al., 2018; Kovanović et al., 2016; Xing Tang, & Pei, 2019).

The qualitative analyses on the forum posts performed by the text classifier revealed that the distribution of social presence varied as the MOOC progressed with different requirements on the assignments. Specifically, in Module 1 and Module 2, where posting links of examples and making comments to others were required, social presence indicators such as *Sharing resources, Using vocatives, complimenting others* and *Offering advice* dominated the discussions, implying that learners were trying to meet the requirements of the assignments by sharing the links of their examples,

commenting on others' posts (evidenced by addressing others using their names), making compliments and giving advice. The data also showed that *Self-disclosure* is particularly high in Module 1, which is expected because of most learners gave a brief self-introduction when they made their first posts in the course, or revealed their own backgrounds when they made comments and offered advice to others. From the affective perspective, one conspicuous finding is that *Positive emotions* dropped by almost 60% from Module 1 to Module 2, while *Negative emotions* remained similar in the first two modules, even though the number of participants dropped by half from Module 1. This trend should alarm the instructors to check these posts that convey negative emotions and figure out possible causes to the change of sentiment among learners (e.g., workload of assignments, complexity of instructional materials, learning curve of new tools etc.). Instructors' timely intervention is necessary to address learners' problems to ensure they receive enough support as they progress to later modules of the MOOC. Ways that instructors might intervene upon noticing such changes in students' sentiments include identifying the individuals who need urgent help, addressing their problems and discussing possible solutions through replying their posts or private messaging.

Similar to the first two modules, *Sharing resources, Using vocatives* and *Complimenting others* were still the most frequent social presence indicators in Module 3, which was mostly assignment-driven. Whereas in Module 4, when providing external links and commenting on others' posts were not required, *Sharing resources* and *Using vocatives* were still relatively high. This implies that learners preferred to use external resources to give examples and illustrate their points. And the high occurrences of *Using vocatives* suggest that the posts were mostly conversational, signaling high level of peer interactions despite the fact that providing more resources and replying to more peers will not result in any bonus points. Interestingly, there was a precipitous drop of

Complimenting others in Module 4 compared to the prior three modules. This can be partly explained by the continuous decrease of participants in the forum. Another reason may be the lack of sharing external resources or one's own work, which was required in Module 1, 2 and 3. In other words, the frequency of *Complimenting others* was proportional to that of *Sharing resources*. Learners tended to praise their peers' assignments when they were required to give comments. More importantly in Module 4, the increased occurrences of *Offering advice, Disagreement/doubts/criticism*, *Self-disclosure* and *Negative emotions* suggest that the discussions became more critical regarding the topic of machine learning in data journalism, which is a rather novel and controversial topic. Learners expressed more negative emotions than positive emotions, communicated their doubts and concerns on the topic, shared their own experience/stories to support their opinions, and gave a great deal of advices to improve the status quo. Compared to posts that mainly expressed compliments and admiration, which are relatively shallow in a cognitive sense, critiquing each other's ideas, communicating disagreements, doubts, exploring alternatives and offering advice seem to be more constructive and signal higher order learning such as applying the course concepts to evaluate others' ideas and work (Sosniak, 1994).

Similar to the Module 1 and 2, Module 5 and 6 also required the provision of links to external resources or one's own work, as well as commenting on others' posts. Again, the high frequencies of *Sharing resources, Using vocatives*, *Complimenting others* and *Expressing gratitude* echo learners' efforts to meet these requirements. Interestingly, *Complimenting others* was the most frequent social presence indicator in both Modules 5 and 6, indicating that learners tended to receive more compliments from peers when they completed and shared the results of more sophisticated tasks (e.g., visualizing data in Module 5 and formulating data stories in Module 6) than simple tasks (e.g., finding data

in Module 1 and preparing data in Module 2). Contrary to Module 4, the frequency of *Positive emotions* exceeded that of *Negative emotions* in both Module 5 and 6, suggesting that learners experienced more joy than confusion and frustration about the topics or in the learning process. In particular, the frequency of *Negative emotions* in Module 5 remained the same with that in Module 4, even with a drop in the number of participants, but was 34% higher than that in Module 6. This may imply that learners were struggling in Module 5, which calls for instructors' timely attention and intervention. Interestingly, *Self-disclosure* in Module 6 was the highest in frequency compared to that in Module 3, 4 and 5, even with fewer participants, meaning that in the context of Module 6, which was about data-driven story-telling, learners' data stories had a lot of personal relevance – typically linked to their professional practice, their country/city of residence, the social issues of their concerns and other highly personal experience. Earlier studies note that self-disclosure is important for social attraction and bonding between individuals. For example, Cutler (1995) explained that "the more one discloses personal information, the more others will reciprocate, and the more individuals know about each other the more likely they are to establish trust, seek support, and thus find satisfaction". According to Lu and Farzan (2015), the level of self-disclosure has a significant positive relationship with learners' subsequent effort in the forum. Those who revealed their personal information and sharing personal stories were more likely to post more and come back to the forums.

Besides meeting the basic requirements of the assignments of sharing the links of one's work/examples, it is interesting to note that there was a consistent high frequency of *Using vocatives* (addressing peers by names) throughout the course. According to Rourke et al. (1999), the researchers who first proposed the social presence framework, group cohesion is exemplified by activities that build and sustain a sense of group

commitment. Vocatives are an important expression of group cohesion because they connote feelings of closeness and association. The teacher immediacy literature has also discovered an empirical connection between addressing students by name and cognitive, affective, and behavioral learning (Christenson & Menzel, 1998; Gorham, 1988; Gorham & Christophel, 1990; Sanders & Wiseman, 1990). Seeking to explain this connection, Kelly and Gorham (1988) found support for a relationship between vocatives and immediacy of recall. Eggins and Slade (1997) support the use of vocatives to facilitate social presence, noting "the use of vocatives would tend to indicate an attempt by the addresser to establish a closer relationship with the addressee".

Besides *Using vocatives* to establish closer ties with peers, the high frequencies of *Complimenting others* and *Expressing gratitude* represent an inclusive and welcoming atmosphere within the forums where students showed respect and appreciation when critiquing each other's work, especially in later stage of the course (Module 5 and 6) as the community became more stable. Within this favorable environment, students also expressed more *Positive emotions* during the processing of learning. Salmon (2000) asserted that students need to pass through a stage of socialization to achieve higher levels of group development and learning. Positive social presence like *Complimenting others* and *Expressing gratitude* help to establish a respectful and friendly environment which is conducive to learning and knowledge construction (Sanders, Ventura, & Dando, 2007).

Going beyond the basic online etiquette, *Offering advice* is an important way to demonstrate one's understanding and application of knowledge. The high frequency of *Offering advice* (especially in Module 4) implies more cognitively valuable comments that demonstrate learners' level of understanding of the course concepts that warrant instructors' attention. There are other social presence indicators that may be useful to

evaluate learners' level of knowledge, such as *Personal opinion/reflection, Disagreement/doubts/criticism,* and *Asking questions.* However, in the context of this course, the consistently higher frequency of *Offering advice* suggests that it is the primary way learners chose to communicate their understanding of the content in the modules.

Learners' affective states during discussions is another dimension may help instructors better target their facilitation by responding to posts that require urgent attention. There are sufficient empirical studies that establish important connections between emotional processes and learning (Atkins, 2002; Pekrun, 2006; Xing et al., 2019). When moving into the online learning environment the emotional dimension becomes even more pertinent (Hughes, Ventura, & Dando, 2007). The increase of *Negative emotions* (especially when it exceeded the occurrence of *Positive emotions* in Module 2 and Module 4) may signal that learners were struggling with learning or expressing negative comments on the topics chosen. This trend warrants instructors' attention and timely intervention.

THE CHANGES IN THE LEARNER NETWORKS IN RELATION TO THE PATTERNS OF SOCIAL PRESENCE

Passive participation

SNA was used to analyze learners' passive and active participation in the forum. During passive participation, which only considers reading posts without creating posts, the declining number of nodes in each module denotes that the learning community became smaller with the attrition of participants as the MOOC progressed. The loss of participants was more severe in the first two modules than in later stage of the course. As the course progressed to Module 4, the community became relatively stable with lower drop-out rates afterwards. When looking at the frequency of learners' latent interaction

with one another through reading posts, which is captured by the number of edges in the network, it appears that the number peaks at Module 1. This is expected because Module 1 had the highest number of participants. After more than half of the learners dropped out from the forum in Module 2, the level of interaction also dropped by more than half, and continued to drop until Module 4. The number of edges slightly increased in Module 4 and Module 5, but resumed to a downward trend in Module 6. This may be explained by the intriguing topics in these two modules that sparked higher reading rate even though the community was shrinking as a whole. The remaining learners showed more commitment to learning by the reading more of their peers' posts.

The calculation of modularity yielded the number of groups within the learner network in each module. The results show a clear trend of decrease in the number of groups as the course progressed (from 10 groups in Module 1 to only 5 groups in Module 6). This is partly due to the decrease of participants over the six modules. The size of groups also became smaller over time. This suggests that the learner network became tighter with a group of devoted learners who persistently read their peers' posts. In Module 4, however, the maximum size of the group grew more than twice larger than that in Module 3, indicating that there were some larger groups with tightly connected individuals who read each other's posts. In other words, there may be some very popular posts/threads that a lot of learners have viewed. For instructors, identifying these posts/threads and bringing them to more audience might be useful to trigger more conversations among learners and create more opportunities for knowledge construction.

The results of passive participation in the forums show an overall declining trend from the first to the last module, with the biggest drop happening in Module 2 in terms of the number of participants and the level of interaction. This finding is consistent with other research (Khalil & Ebner, 2016; Liu et al., 2019; Qu & Chen, 2015; Tseng et al.,

2016) showing a rapid decline of the number of participants and less learning activities occurred after the first one or two modules. This result suggests that the instructional design and support are critical in the first two modules to retain participants in MOOCs. Possible strategies include using intriguing and easily digestible instructional materials, setting up discussions around important topics, providing more instructor support etc. This finding is especially important in light of the high dropout rates found in the MOOC literature (Fidalgo-Blanco et al., 2016; Gütl et al., 2014; Joo, So, & Kim, 2018).

Some of the studies discouraged passive participation and considered lurkers to be free-riders (Kollock & Smith, 1996; Morris & Ogan, 1996; Rheingold, 2000). The sustainability of an online community needs new content and timely interactions, but passive participants contribute little value to the community and may impair the vitality of the community (van Mierlo, 2014). On the contrary, other studies argued that lurking is not only normal but also is an active, participative and valuable form of online behavior (Edelmann, 2013). Many passive participants thought of themselves as a part of the community, and lurking was their preferred way to engage in learning within the community (Nonnecke et al., 2006), because they felt they were learning just as much or more from reading others' posts than from writing their own (Beaudoin, 2002). Lave and Wenger (1999) considered lurking behavior in a community of practice as a form of cognitive apprenticeship, whereby novices learn by observing experts within a social context. As novice learners accumulate more knowledge and skills, their participation becomes more visible, until they develop mastery of the shared repertoire of the learning community. Therefore, even though passive participation is less visible, these lurking behaviors equally result in knowledge exchange and contribute as much as active participation (Zhang & Storck, 2001).

Active participation

Similar to passive participation, the number of participants in the network gradually declined as the MOOC progressed over time. The most drastic decrease of participants occurred in Module 3. As the course moved on to Module 4, the declining trend started to slow down. And there was an uptick of interaction in Module 5. This interruption of the downward trend can be explained by the intriguing topics of Module 4 (machine learning in data journalism) and Module 5 (visualizing data). As the learner network shrunk in every module, the network became gradually tighter, implying that the general level of connections among the learners grew higher by module with a group of persistent and devoted learners. These findings echo previous literature studying the patterns of peer interaction in online learning communities, in which core-periphery participation structures and the prevalence of weak ties are common (Aviv, Erlich, & Ravid, 2007; Butts, 2008; Jones, Ferreday, & Hodgson, 2008).

Even though the network gradually grew denser, there were a large number of isolated individuals who only posted their responses to the forum questions but not interacting with any peers in Module 3 and 4. The low edges also imply that a lot of learners who posted their responses to the assignments did not receive any replies from peers. This can be partly explained by the fact that commenting on others' posts was not required in these two modules. However, the absence of these isolated nodes and the relatively higher network density in the passive participation networks in these two modules suggest that those posts did receive views from peers although without explicit responses. This finding mirrors Kellogg, Booth, and Oliver's (2014) study that non-mutual interactions and sparse network are common in MOOC discussion forums. More radically, as reported in Wasko, Teigland, and Faraj's (2009) study, half of the network

consisted of "outsiders" who did not receive responses, and "seekers" who received responses but did not reciprocate, thus resulting in low network density.

THE RELATIONSHIP BETWEEN LEARNERS' CENTRALITY AND SOCIAL PRESENCE

The correlations between social presence and learners' centrality

The correlation results between the 13 social presence indicators and the four centrality measures show that, *Expressing agreement*, *Expressing gratitude,* and *Disagreement/doubts/criticism* had strong to moderate correlations with all four centrality measures. This implies that learners who expressed more agreement and gratitude, and voiced out their disagreement and doubts were more likely to reach a central position in the learner network. Besides these three social presence indicators, *Negative emotions* was found to have a moderate correlation with in-degree, while *Referencing others* had a moderate correlation with both in-degree and Eigen centrality. This suggests that expressing negative emotions in the forum may result in more replies from peers; whereas quoting and referencing others' ideas were more likely to trigger others' responses and help a learner establish social ties with more well-connected peers (those who had higher in-degree). However, these two social presence indicators were not related to closeness and betweenness centrality in any way.

The positive correlations between *Expressing agreement*, *Disagreement/doubts/criticism* and the four network centrality measures suggest that taking stance (either agreement or disagreement) on a topic or on others' opinions may lead to more central positions in the learner network. Though it not hard to understand that seconding others' opinions and expressing gratitude are beneficial to establish peer rapport, it is interesting that *Disagreement/doubts/criticism* may also achieve similar results. In social learning contexts, it is not uncommon to see the emergence of

disagreement, typically when a learner raises doubts and concerns when discussing a topic or critiquing the work of others. In doing so, they may draw attention to the negative aspects of a topic or risk offending others by pointing out the problems, but the positive correlation between *Disagreement/doubts/criticism* and the four network centrality measures suggests that it is conducive in establishing rapport with peers and helps a learner to earn a more favorable position in the learner network. From a cognitive sense, it is important to bring up conflicting viewpoints because dissenting information may disrupt one's existing cognitive framework and create a state of disequilibrium, triggering a learner to reflect on his or her prior knowledge and make adjustments to accommodate the new information. This process enables higher order thinking and is crucial for knowledge construction, according to the view of constructivist learning (Vygotsky, 1980; Piaget, 2013). Therefore, learners should not be afraid to be more critical in their posts, while instructors should encourage dissenting voices as long as learners can provide sound justifications to support their ideas.

Another interesting finding is that *Negative emotions* positively correlated with learners' in-degree, which means expressing more negative emotions in the forums may result in more responses from peers. Due to the diverse backgrounds of learners in MOOC contexts, it is natural that some of them experienced struggles because of the lack of prerequisite knowledge, resulting in the emergence of negative emotions such as feeling challenged and confused. The fact that these negative emotions sparked more responses from others may be because many learners encountered the same problems and experienced the same feelings. Expressing negative emotions is necessary because voicing out confusion, frustration and complaints may help the instructors locate the struggling learners and the difficulties in the course content. Over the last five years, there have been a lot of research efforts focusing on identifying learners' emotions,

confusion and help-seeking behaviors by analyzing their forum posts (Agrawal, Venkatraman, Leonard, & Paepcke, 2015; Almatrafi, Johri, & Rangwala, 2018; Hecking, Hoppe, & Harrer, 2015; Chandrasekaran, Kan, Tan, & Ragupathi, 2015). These studies developed new techniques attempting to accurately detect learners' struggles during the learning process. As these new techniques gradually become available in the MOOC forums, learners' explicit expression of negative emotions when encountering difficulties will help alert the instructors to step in and provide timely support. Previous studies on MOOCs provided ample evidence that struggling learners were most vulnerable to lose motivation and drop out due to delayed or lack of timely support (Agrawal et al., 2015; Wen et al., 2014; Xing, Chen, Stein, & Marcinkowski, 2016). However, in online learning contexts with massive number of learners, it is hard for the instructors to identify struggling learners if the learner did not make explicit their problems or ask for help.

Predicting Network Centrality from Social Presence and Posting Behaviors

In the hierarchical regression models that predict in-degree, the base model (without posting behaviors as predictors) showed that the seven social presence indicators (*Complimenting others, Disagreement/doubts/criticism, Expressing agreement, Expressing gratitude, Negative emotions,* Offering advice*, Referencing others*) contributed significantly to the regression model, accounting for 11.3% of the variation in in-degree. While in the prediction of Eigen centrality, *Asking questions* and *Using vocatives* were two more additional significant predictors in the base model. Introducing out-degree in the prediction explained an additional 67.3% of the variations in in-degree, and 51.4% of the variations in Eigen centrality. This suggests that the total number of the posts a learner created in the forums contributed significantly to the number of replies he or she received, as well as the quality of his or her connections in the learner network.

This echoes an earlier study by Yusof and Rahman (2009) that learners who contributed more to the forum also had more friendly relations with peers and assumed important roles in delivering information to the learning community. By contrast, including the *average length of post* as a predictor to the regression models added very low predicting power to in-degree, and zero additional power to Eigen centrality. This implies that the length of posts had very low or no effect on the number of replies one receives or the quality of connections one made. However, based on the findings of this study that certain social presence indicators are correlated with higher centrality, crafting longer and more in-depth posts may be beneficial to trigger more responses among peers. However, it is also possible that learners who were extrinsically motivated to participate in the forum (those who only complete the forum assignments to obtain the course certificate) may choose shorter posts to comment on, in order to finish the tasks with less investment of time and efforts. Finally, the addition of *posting day* as a predictor to the regression model added slightly more prediction power to the model (0.7% to in-degree, and 3.4% to Eigen centrality). The effects were small but still statistically significant. This indicates that the day of posting still mattered even though its impact was relatively small compared to social presence and out-degree. While the effect of timing of posting varied in different studies (Jaech et al., 2015; Lampe & Resnick, 2004; Mazzolini & Maddison, 2007), this study indicates that, the earlier one post his or her assignments/comments, the more replies he or she may receive from peers, and in the meantime more likely to attract the responses from well-connected peers. This is expected because earlier posts are more likely to be read and commented by audience when they have limited choice of posts to read and reply to early on.

In the prediction of closeness centrality, the base model showed that *Asking questions, Complimenting others, Expressing agreement, Expressing gratitude,* and

Using vocatives contributed significantly to the model, accounting for 11.5% of the variations in closeness centrality. These five social presence indicators are also the significant predictors of Eigen centrality in the base model. However, unlike the prediction of in-degree and Eigen centrality, out-degree only explained an additional 6.6% of the variations in closeness centrality. This implies that the total number of posts one created in the forum did not contribute to the distance between this individual and all other peers in the learner network. This result suggests that the learners in this MOOC, even for active learners who created more posts, were interacting with a limited number of participants (or the same group of peers) throughout the course instead of reaching out to a wider range of participants, thus failed to shorten the distance with others in the network. This mirrors the findings of Vaquero and Cebrian's (2013) study that discovered a phenomenon of group exclusivity in online interactions: more frequent and intense social interactions were hosted within a stable group of learners (typically mediated by persistent interactions among high performing students). New participants' attempts to engage in the conversations in those established groups often failed to produce reciprocity, in which low performing students were selectively excluded. Failure to engage in the "rich club" eventually decreased low performers' communication activities towards the end of the course. This exclusivity of group interactions may be detrimental to learners' motivation to engage in the conversations with new peers. The lack of reciprocity may result in higher dropout rate or peripheral participation, and ultimately decrease the chance of knowledge construction within the learning community (Lave & Wenger, 1991; Nistor, Dascalu, Serafin, & Trausan-Matu, 2018).

In the regression models that predict betweenness centrality, the base model had four significant social presence predictors, namely *Complimenting others, Expressing agreement, Negative emotions* and *Offering advice*. It is worth noting that these four

social presence indicators also significantly predict in-degree in the base model. However, they only accounted for 3.8% of the variations in betweenness centrality. Compared with the prediction of the other three centrality measures, the impact of social presence is relatively small on betweenness centrality. But introducing out-degree in the model significantly improved its predicting power by adding an additional 54.5% of the variations in betweenness centrality. This is similar to the models of in-degree and Eigen centrality, with out-degree playing a major role in the prediction. This suggests that contributing more posts to the forums increased the likelihood of being the "bridges" between peers, thus having more influence on the information flow in the network. Identifying and interacting with these individuals may help instructors broadcast meaningful information to the learning community more efficiently and effectively (Van der Hulst, 2009). Finally, similar to the case of other three centrality measures, including *average length of post* and *posting day* to the prediction added minimal or zero predicting power to betweenness centrality.

THE CORRELATION BETWEEN LEARNERS' CENTRALITY AND LEARNING OUTCOME

The correlation analyses between learners' centrality and learning outcomes revealed that certificate status was strongly correlated with in-degree, closeness centrality, betweenness centrality, and moderately correlated with Eigen centrality. This suggests that learners who received more replies in the forums, had more high-quality connections, shorter distance with others, and served more frequently as "bridges" between others were more likely to receive the course certificate in the end. While prior literature had inconsistent findings regarding what centrality measures are linked to course completion status (Houston et al., 2017; Jiang et al., 2014; Joksimović et al., 2016), this study provides empirical evidence of a positive correlation between learners'

centrality and their course completion status. Based on the results of hierarchical regressions that out-degree being a major predictor of centrality measures, it is reasonable to assume that in order to obtain central positions in the network, learners need to contribute more posts and engage in more conversations than others in the forums. These active posters, according to Lave and Wenger (1991), can be classified as central participants in a learning community. When learners participate actively and consistently in the forum, they gradually develop their social identity as core members in the community. Lave and Wenger (1991) argued that this social identify is inseparably intertwined with the cognitive components within a learning community. Specifically, central participants assume more responsibility and perform more difficult tasks than peripheral members; therefore, their identity is that of an expert, which is both a cognitive attribute and a socially negotiated status (Lave & Wenger, 1991). The social identity of a "central learner" as a leader/expert may trigger a sense of responsibility and higher motivation to perform better than others, thus ultimately completing the MOOC with a certificate.

By contrast, learners' centrality appeared to have no association with their perceived learning from general, cognitive, affective and behavioral perspectives, as indicated by the correlation results. This suggests that learners' centrality in the network is not linked to their perceived learning in any way. This finding is contradictory to prior literature that shows more centrally situated learners tend to get higher final grades (Romero, López, Luna, & Ventura, 2013) and more desirable cognitive learning outcome (Russo & Koesten, 2005). The reason for this could be, there are rich learning resources in the MOOC besides the discussion forums, such as videos, require and optional readings, quizzes etc. Learners may use those course components more heavily than discussion forums to obtain the knowledge and skills they chose to learn. An earlier study

by Liu, Zou, Shi, Pan, and Li (2019) found that of all course components in a MOOC, the discussion forum was ranked the least favorite by learners. Participants' qualitative responses revealed that they did not consider discussion forums to be time well spent, due to the course's massive nature, lack of meaningful interactions with the instructor, and poor feedback from peers. These negative impressions towards discussion forum may inhibit learners' motivation to participate. As Jung and Lee (2018) pointed out, perceived usefulness has a direct effect on learning engagement. Therefore, learners may not necessarily be the most central and active participants in the forum, but they can still learn the knowledge and skills they desire due to the rich choices of available learning materials provided by the MOOC.

Finally, when it comes to satisfaction, it is found that only Eigen centrality was significantly correlated with learners' satisfaction, which implies that if a learner interacts with more influential peers in the forums, he or she is more likely to feel satisfied with the learning experience in the MOOC. This is expected because the opportunities to converse with more central participants to constantly negotiate meanings increase the likelihood of knowledge construction. Collins, Brown and Newman's (1988) cognitive apprenticeship theory posits people learn from one another, through observation, imitation and modeling. The active and well-connected individuals in the network can be a great source for learning since they are more likely to reciprocate the interactions, thus more chances to provide more advice/feedback/resources to help others master the knowledge and skills they intend to learn. It is reasonable to assume that those who benefited from this process experienced higher level of satisfaction in the course.

CONCLUSION

Discussion forums are widely provided in MOOCs for learners to interact and exchange learning support. Regularly engaging in the forums to share one's understanding of the course content, interact with others, negotiate meaning and develop higher order thinking can be an important form of MOOC learning. To further understand learners' engagement in MOOC discussion forums, this study adopted a mixed-method approach to examine learners' participation patterns and social presence throughout the course, and the relationship between their social presence, centrality in the learner network, and learning outcomes.

The qualitative analyses on the forum posts performed by the text classifier revealed that the distribution of social presence varied as the MOOC progressed with different requirements on the assignments. Specifically, in the first and last two modules (Module 1, 2, 5, 6), where posting links of examples/resources/assignments and making comments to others were required, social presence indicators such as *Sharing resources, Using vocatives,* and *Complimenting others* dominated the discussions, implying that learners were trying to meet the requirements of the assignments by sharing urls, commenting on others' posts (evidenced by addressing others using their names) and making compliments. Whereas in Module 3 and 4, when providing external links and commenting on others' posts were not required, *Sharing resources* and *Using vocatives* were still relatively high. This implies that learners preferred to use external resources to give examples and illustrate their points. And the high occurrences of *Using vocatives* suggest that the posts were mostly conversational, signaling high level of peer interactions. Going beyond meeting the basic requirements, the relatively high occurrences of *Offering advice* in Module 3 and 4 imply that learners posted more cognitively valuable comments that demonstrate their level of understanding of the

course concepts. The fluctuating frequencies of *Positive emotions* and *Negative emotions* reflect the changes of learners' affective states in the forums, which corresponds to the changes of topics and varied levels of complexity of the tasks in each module.

The changes of passive and active participation are somewhat similar across the six modules. The loss of participants is more severe during the early stage of the course. As the course progressed to Module 3 and 4, the community became relatively stable with lower drop-out rates afterwards. The level of peer interactions (both reading and posting) peaked at Module 1 with the largest number of participants, but with the lowest network density. The number of edges slightly increased in Module 4 and Module 5, but resumed to a downward trend in Module 6. This may be explained by the intriguing topics in these two modules that sparked higher reading and posting rate even though the network was shrinking as a whole.

The correlation results between the 13 social presence indicators and the four centrality measures show that, *Expressing agreement*, *Expressing gratitude,* and *Disagreement/doubts/criticism* had strong to moderate correlations with all four centrality measures. This implies that learners who expressed more agreement and gratitude, and voiced out their disagreement and doubts were more likely to reach a central position in the learner network. Besides these three social presence indicators, *Negative emotions* was found to have a moderate correlation with in-degree, while *Referencing others* had a moderate correlation with both in-degree and Eigen centrality. This suggests that expressing negative emotions in the forum may result in more replies from peers; whereas quoting and referencing others' ideas were more likely to trigger others' responses and help a learner establish social ties with more well-connected peers (those who had higher in-degree). However, these two social presence indicators were not related to closeness and betweenness centrality in any way.

In the prediction of learners' centrality using only social presence indicators, the results show that *Complimenting others* and *Expressing agreement* significantly predict all four centrality measures. Additionally, *Negative emotions* and *Offering advice* were two important predictors of in-degree, Eigen centrality and betweenness centrality, while *Asking questions* and *Using vocatives* were important in the prediction of both Eigen centrality and closeness centrality. However, after including posting behaviors in the regression models, out-degree became the most dominant predictor of all four centrality measures, whereas the effects of the average length of posts and the day of posting were marginal.

The correlation analyses between learners' centrality and learning outcomes revealed that certificate status was strongly correlated with in-degree, closeness centrality, betweenness centrality, and moderately correlated with Eigen centrality. This suggests that central participants of the community were more likely to complete the course and received the certificate in the end. By contrast, learners' centrality appeared to have no association with their perceived learning from general, cognitive, affective and behavioral perspectives, as indicated by the correlation results. This suggests that learners' perceived learning was not necessarily linked to their engagement in the forums and their positions in the learner network. When it comes to satisfaction, it is found that only Eigen centrality was significantly correlated with learners' satisfaction, which implies that if a learner interacts with more influential peers in the forums, he or she is more likely to feel satisfied with the learning experience in the MOOC.

IMPLICATIONS

As a theoretical contribution, this study fills the gap of understanding the relationship between social presence, learners' network centrality and learning outcomes.

It provides a critical ground for studying content-related interaction and learning community in MOOC forums. The findings will inform MOOC learners in terms of how to strategically present themselves in the discussion forums to increase the possibilities of more peer interactions and achieve productive learning outcomes. While for MOOC instructors, this study will potentially inform them how to effectively mediate the discussions and improve learner engagement as a facilitator. As a methodological contribution, this study proposes a theoretically grounded computational linguistics model based on the chosen framework of social presence to automate the analysis of students' forums posts. Specifically, the findings of this study provide the following implications for learning and teaching in MOOC contexts:

For learners, contributing more posts to the forums is essential for obtaining more central positions in the learner network. When creating posts, it is important to strategize the content of the post by taking account of the impact of social presence. In particular, taking a stance by expressing agreement or disagreement, addressing negative feelings, bringing up the problems/caveats or even criticism can be conducive to attract more attention from others. More importantly, explicitly expressing negative emotions in the forum when encountering difficulties may help alert the instructors to step in and provide timely support. Besides that, it is also crucial to give sincere compliments to others when they have done a job on a task. After receiving compliments or feedback/advice from others, a timely response to express gratitude is helpful for making one more visible to the whole community. Beyond basic online etiquette, providing valuable advice to help peers improve their assignments is also beneficial for one's position in the network. In order to develop closer social ties with others in the community, or attracting the attention of more influential peers, it is essential to avoid staying within the same circle of peers throughout the course. Making an effort to reach out to a variety of peers

increases the chance to enrich one's knowledge and obtain a more favorable position in the network. It is also important to keep in mind that asking questions and addressing peers using their names are useful to develop more social ties with peers, especially influential peers. Additionally, quoting and referencing others' ideas are also helpful to trigger others' responses and help a learner establish social ties with more well-connected peers. Even though the average length of posts and the timing of posting have smaller impact on one's network position compared to social presence, it is always beneficial to draft more in-depth posts that clearly outline one's understanding or arguments, and post in a timely manner to increase the chance to reach a broader audience.

From an instructor's perspective, the findings from this study provide some meaningful implications in facilitating forum discussions. Firstly, it is important for instructors to monitor the affective states of students in each module. For example, in the context of this study, the occurrence of *Negative emotions* exceeded *Positive emotions* in Module 2 and Module 4, which signals that learners were struggling with the learning processes or expressing negative comments on the topics of that module. Expressing negative emotions is necessary because voicing out confusion, frustration and complaints helps the instructors identify struggling learners. Also, when these negative comments accumulate, it may suggest that many learners are struggling with the similar problems and experiencing the same feelings. This calls for instructors' immediate attention and timely intervention. Secondly, it is meaningful to encourage and facilitate learners to formulate arguments or taking a stance on a given topic, since the expressions of both agreement or disagreement are more likely to trigger peer interaction. For example, asking learners to debate about a novel or controversial topic may be a good strategy to drive up learner engagement in the forum. In the context of the MOOC in this study, the discussion of data preparation in Module 2 could have been changed to a debate format

(e.g., asking learners to evaluate and debate about different approaches of data cleansing) in order to foster more student interaction. In the meantime, instructors should provide a safe and friendly environment for open discussions, and make sure that all opinions are welcomed as long as one can provide sound justifications to support his or her ideas. When learners feel more trust in the instructors and peers in the community, they might be more willing to voice out disagreements, concerns or even criticism, which are all valuable elements to stimulate learner participation. Furthermore, paying attention to the changes of network structure and identifying important posts in the learning community may help bring more meaningful content to learners and increase the overall engagement in discussion forum. For example, the existence of larger groups within a network may suggest that there are some intriguing topics/content that attract the comments from a lot of participants. For instructors, is it meaningful to identify these popular topics/content and bring them to more learners (e.g., pin the discussions of those topic to the top of the forum). Instructors' stepping in and participating in the discussion of those topics/content may also increase the overall learner engagement in the forum. Last but not least, it is beneficial to identify active individuals in the forums since they are more likely to serve as the "bridges" between peers and have more influence on the information flow in the network. Interacting with these individuals may help instructors broadcast meaningful information to the learning community more efficiently and effectively.

The findings of this study also provide meaningful insights into the system design of discussion forums in online learning settings. First, automatic content analysis algorithms should be used to enhance the function of the forum based on learners' social presence. The results of such analysis could be visualized and presented on student or instructor dashboards. As suggested in this study, the high frequencies of certain social presence indicators may reveal plenty of useful information to inform the instructors'

pedagogical decisions. For example, higher levels of negative emotions or disagreement/doubts/criticism or more questions asked in the forums can alert instructors/ timely intervention. Second, since both passive and active participation (reading and posting) are equally important in learning, it is meaningful to inform the learners about their level of passive and active participation the attention. For example, present the number of views/replies of one receives for each of his or her post. Furthermore, it may be helpful to visualize and inform learners about their centrality in the learner network. In this way, learners can have a clearer understanding of their level of engagement in the discussion forum compared to their peers, which may motivate them to catch up when they are falling behind, or to give them a sense of achievement when they are ahead of others.

LIMITATIONS AND FUTURE WORK

This study examines the learner engagement in MOOC discussion forum from the perspectives of social presence, the structure of learner network and learners' centrality. There are four limitations of this study. Firstly, due to the unique context of this study, the results may not be generalizable. Further studies are needed to investigate the impact of social presence using datasets from MOOCs of other topics to test the reliability and validity of our findings. Beyond the MOOC context, future studies should also look into credit-based online learning course in higher education settings, given the rapid transition from traditional classroom to online education, especially during public health crisis. The different structure of learning communities and social dynamics may yield meaningful insights that are highly relevant and valuable to facilitate online learning and teaching. Secondly, the sampled population may be biased because survey responses only came from post creators. And survey respondents tend to be more active and responsive

learners by nature. There is no survey data from learners who observed the forums throughout the six modules but never posted. Plus, the small sample size of survey respondents affects the reliability of the findings, particularly the correlation between learners' network centrality and their perceived learning and satisfaction. Furthermore, the analysis of learners' social presence mainly relied on the automatic classification performed by the text classification model, the accuracy of which still needs to be improved, especially in MOOC settings where learners have vastly different styles of written communication given their diverse cultural/professional backgrounds and varied levels of English proficiency. Future studies should aim at constructing larger training sets to build more robust content analysis models, improving the overall performance, and undertaking content analysis on a deeper level. For example, unpacking negative/positive emotions or agreement/disagreement to obtain more in-depth knowledge of learners' posts. Additionally, considering the limitation of automatic content analysis, this study only focuses on the analysis of social presence. However, under of premise of CoI, teaching presence and cognitive presence are equally important for cultivating successful knowledge communities. Future studies should incorporate the examination of teaching presence and cognitive presence to understand how these three elements interact to impact students' learning outcome. Finally, methodologically, this study used SNA and content analysis to examine the role of social presence in learners' engagement. This is among the first efforts to combine network and content analysis of online learning forums. Future studies may explore other integrative methods that combine network and semantic analysis, such as cohesion network analysis (Dascalu, McNamara, Trausan-Matu, & Allen, 2018) and epistemic network analysis (Shaffer et al., 2009), to model the interactions between learners from both social and cognitive perspectives.

Reference

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: Addressing confusion in MOOC discussion forums by recommending instructional video clips. *Proceedings of the 8th international conference on education data mining* (pp. 297–304). New York, NY, USA: ACM.

Akyol, Z., & Garrison, D. R. (2008). The development of a community of inquiry over time in an online course: Understanding the progression and integration of social, cognitive and teaching presence. *Journal of Asynchronous Learner Networks*, *12*, 3-22.

Akyol, Z., Vaughan, N., & Garrison, D. R. (2011). The impact of course duration on the development of a community of inquiry. *Interactive Learning Environments*, *19*(3), 231-246.

Almatrafi, O., Johri, A., & Rangwala, H. (2018). Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education*, *118*, 1-9.

Al-Rahmi, W. M., Alias, N., Othman, M. S., Marin, V. I., & Tur, G. (2018). A model of factors affecting learning performance through the use of social media in Malaysian higher education. *Computers & Education*, *121*, 59-72.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014, April). Engaging with massive online courses. *Proceedings of the 23rd international conference on World wide web* (pp. 687-698). ACM.

Annand, D. (2011). Social presence within the community of inquiry framework. *The International Review of Research in Open and Distributed Learning*, *12*(5), 40-56.

Arbaugh, J. (2008). Does the community of inquiry framework predict outcomes in online MBA courses? *International Review of Research in Open and Distance Learning, 9*(2), 1–21.

Arbaugh, J. B., Bangert, A., & Cleveland-Innes, M. (2010). Subject matter effects and the community of inquiry (CoI) framework: An exploratory study. *The Internet and Higher Education*, *13*(1-2), 37-44.

Arbaugh, J. B., & Hwang, A. (2006). Does "teaching presence" exist in online MBA courses? *The Internet and Higher Education*, *9*(1), 9-21.

Arthur, C. (2006). *What is the 1% rule?* The Guardian. UK. Retrieved from: https://www.theguardian.com/technology/2006/jul/20/guardianweeklytechnologysection2#:~:text=It's%20an%20emerging%20rule%20of,89%20will%20just%20view%20it.

Atapattu, T., Falkner, K., & Tarmazdi, H. (2016). Topic-wise classification of MOOC discussions: A visual analytics approach. *Proceedings of the 9th International conference on Educational Data Mining*, Raleigh, NC, USA.

Atkins, J. (2002). The emotional dimension of learning. *Learning in Health and Social Care*, *1*(1), 61-62.

Aviv, R., Erlich, Z., & Ravid, G. (2007). Randomness and clustering of responses in online learning networks. *Proceedings of the Communication, Internet, and Information Technology 2007*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.9151&rep=rep1&type=pdf

Baggaley, J. (2013). MOOC rampant. *Distance Education*, *34*(3), 368-378.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist*, *44*(9), 1175.

Barak, M., Watted, A., & Haick, H. (2016). Motivation to learn in massive open online courses: Examining aspects of language and social engagement. *Computers & Education*, *94*, 49-60.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, *8*(2009), 361-362.

Beaudoin, M. F. (2002). Learning or lurking: Tracking the ‘‘invisible’’ online student. *The Internet and Higher Education, 5*(2), 147–155.

Bergner, Y., Kerr, D., & Pritchard, D. E. (2015). Methodological challenges in the analysis of MOOC data for exploring the relationship between discussion forum views and learning outcomes. In J. G. Boticario, & O. C. Santos (Eds.), *Proceedings of the 8th international conference on educational data mining* (pp. 234–241). Madrid, Spain: International Educational Data Mining Society. Retrieved from http://www.educationaldatamining.org/EDM2015/proceedings/full234-241.pdf

Bischoff, A. (2000). The elements of effective online teaching: Overcoming the barriers to success. *The online teaching guide: A handbook of attitudes, strategies, and techniques for the virtual classroom*, 57-72.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn* (Vol. 11). Washington, DC: National academy press.

Brinton, C. G., & Chiang, M. (2015, April). MOOC performance prediction via clickstream data and social learning networks. *Proceedings of 2015 IEEE conference on computer communications (INFOCOM)* (pp. 2299-2307). IEEE.

Butts, C. T. (2008). Social network analysis: A methodological introduction. *Asian Journal of Social Psychology, 11*(1), 13–41. doi:10.1111/j.1467-839X.2007.00241.x

Caspi, A., & Blau, I. (2008). Social presence in online discussion groups: Testing three conceptions and their relations to perceived learning. *Social Psychology of Education*, *11*(3), 323-346.

Chandrasekaran, M. K., Kan, M. Y., Tan, B. C., & Ragupathi, K. (2015). Learning instructor intervention from MOOC forums: Early results and issues. *arXiv preprint arXiv:1504.07206*.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243.

Christenson, L., & Menzel, K. (1998). The linear relationship between student reports of teacher immediacy behaviors and perceptions of state motivation, and of cognitive, affective and behavioral learning. *Communication Education, 47*, 82-90.

Christophel, D. M. (1990). The relationships among teacher immediacy behaviors, student motivation, and learning. *Communication Education, 39*, 323-34

Chuang, I., & Ho, A. D. (2016). *HarvardX and MITX: Four years of open online courses, Fall 2012-Summer 2016*. Retrieved from SSRN website: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2889436

Chung, C. K., & Pennebaker, J. W. (2014). Using computerized text analysis to track social processes. *The Oxford Handbook of Language and Social Psychology*, 219-230.

Cobb, S. C. (2011). Social Presence, Satisfaction, and Perceived Learning of RN-to-BSN Students in Web-Based Nursing Courses. *Nursing Education Perspectives*, *32*(2), 115.

Coffrin, C., Corrin, L., de Barba, P., & Kennedy, G. (2014, March). Visualizing patterns of student engagement and performance in MOOCs. *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 83-92). ACM.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic press.

Cohen, J., N. Onunaku, S. Clothier, and J. Poppe. (2005). *Helping young children succeed: Strategies to promote early childhood social and emotional development.* Paper Presented at the National Conference of State Legislatures. Washington, DC. Retrieved from: http://www.zerotothree.org/site/DocServer/helping young children succeed final.pdf?docID=1725

Collins, A., Brown, J. S., & Newman, S. E. (1988). Cognitive apprenticeship. *Thinking: The Journal of Philosophy for Children*, *8*(1), 2-10.

Creswell, J. W. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative approaches to research*, (2nd Ed.), Merrill/Pearson Education, Upper Saddle River, NJ.

Croft, N., Dalton, A., & Grant, M. (2010). Overcoming isolation in distance learning: Building a learning community through time and space. *Journal for Education in the Built Environment*, *5*(1), 27-64.

Coakes, S. (2005), *SPSS: Analysis without Anguish Using SPSS v12*, Wiley, Brisbane.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi: 10.1007/BF02310555

Cunningham, J. M. (2015). Mechanizing people and pedagogy: Establishing social presence in the online classroom. *Online Learning*, *19*(3), 34-47.

Cutler, R. (1995). Distributed presence and community in Cyberspace. *Interpersonal Computing and Technology: An electronic Journal for the 21st Century, 3*(2), 12-32.

Dabbagh, N., & Kitsantas, A. (2012). Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. *The Internet and Higher Education*, *15*(1), 3-8.

Damm, C. A. (2016). Applying a community of inquiry instrument to measure student engagement in large online courses. *Current Issues in Emerging eLearning, 3*(1).

Dascalu, M., McNamara, D. S., Trausan-Matu, S., & Allen, L. K. (2018). Cohesion network analysis of CSCL participation. *Behavior Research Methods*, *50*(2), 604-619.

DeBoer, J., Stump, G. S., Seaton, D., & Breslow, L. (2013, June). Diversity in MOOC students' backgrounds and behaviors in relationship to performance in 6.002x. *Proceedings of the sixth learning international networks consortium conference* (Vol. 4, pp. 16-19). Retrieved from: https://www.researchgate.net/profile/Daniel_Seaton/publication/237092327_Diversity_in_MOOC_Students'_Backgrounds_and_Behaviors_in_Relationship_to_Performance_in_6002x/links/59506424a6fdccebfa69f405/Diversity-in-MOOC-Students-Backgrounds-and-Behaviors-in-Relationship-to-Performance-in-6002x.pdf

De Laat, M., Lally, V., Lipponen, L., & Simons, R. J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning*, *2*(1), 87-103.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dillahunt, T., Wang, Z., & Teasley, S. D. (2014). Democratizing higher education: Exploring MOOC use among those who cannot afford a formal education. *International Review of Research in Open and Distributed Learning*, *15*(5), 177-196.

Dowell, N. M., Cade, W. L., Tausczik, Y., Pennebaker, J., & Graesser, A. C. (2014). What Works: Creating Adaptive and Intelligent Systems for Collaborative Learning Support. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 124–133). Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-07221-0_15

Dowell, N. M., & Graesser, A. C. (2014). Modeling learners' cognitive, affective, and social processes through language and discourse. *Journal of Learning Analytics*, *1*(3), 183-186.

Dowell, N. M., Skrypnyk, O., Joksimović, S., Graesser, A. C., Dawson, S., Gašević, D., ...Kovanović, V. (2015). Modeling learners' social centrality and performance through language and discourse. *Proceedings of the 8th international conference on educational data mining* (pp. 205–257). Madrid: IEDMS.

Du, Y. (2006, November). Modeling the behavior of lurkers in online communities using intentional agents. *Proceedings of 2006 International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'06)* (pp. 60-60). IEEE.

Ebner, M., Holzinger, A., & Catarci, T. (2005). Lurking: An underestimated human-computer phenomenon. *IEEE MultiMedia*, *12*(4), 70-75.

Edelmann, N. (2013). Reviewing the definitions of ‘‘Lurkers’’ and some implications for online research. *Cyber psychology, Behavior, and Social Networking, 16*(9), 645–649.

Eggins, S., & Slade, D. (1997). *Analyzing casual conversation*. Washington, DC: Cassell.

Fan, Y. W., Wu, C. C., & Chiang, L. C. (2009). Knowledge sharing in virtual community: The comparison between contributors and lurkers. *Proceedings of the 9th International Conference on Electronic Business*. (pp. 662–668). Macau.

Fedus, W., Goodfellow, I., & Dai, A. M. (2018). MaskGAN: better text generation via filling in the_. *Proceedings of the Sixth International Conference on Learning Representations (ICLR)*. Retrieved from: https://arxiv.org/pdf/1801.07736.pdf?source=post_page

Ferguson, R., & Clow, D. (2015, March). Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 51-58). ACM.

Fidalgo-Blanco, Á., Sein-Echaluce, M. L., & García-Peñalvo, F. J. (2016). From massive access to cooperation: lessons learned and proven results of a hybrid xMOOC/cMOOC pedagogical approach to MOOCs. *International Journal of Educational Technology in Higher Education*, *13*(1), 24.

García-Peñalvo, F. J., Fidalgo-Blanco, Á., & Sein-Echaluce, M. L. (2018). An adaptive hybrid MOOC model: Disrupting the MOOC concept in higher education. *Telematics and Informatics*, *35*(4), 1018-1030.

Garrison, D. R. (2007). Online community of inquiry review: Social, cognitive, and teaching presence issues. *Journal of Asynchronous Learner Networks*, *11*(1), 61-72.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, *15*(1), 7-23.

Garrison, D. R., Anderson, T., & Archer, W. (2003). A theory of critical inquiry in online distance education. *Handbook of Distance Education*, *1*, 113-127.

Garrison, D. R., & Arbaugh, J. B. (2007). Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and Higher Education*, *10*(3), 157-172.

Garrison, D. R., & Cleveland-Innes, M. (2005). Facilitating cognitive presence in online learning: Interaction is not enough. *The American Journal of Distance Education*, *19*(3), 133-148.

Garrison, D. R., Cleveland-Innes, M., & Fung, T. S. (2010). Exploring causal relationships among teaching, cognitive and social presence: Student perceptions of the community of inquiry framework. *The Internet and Higher Education, 13*(1-2), 31-36.

Gillani, N., & Eynon, R. (2014). Communication patterns in massively open online courses. *The Internet and Higher Education*, *23*, 18-26.

Gillani, N., Eynon, R., Osborne, M., Hjorth, I., & Roberts, S. (2014). Communication communities in MOOCs. *arXiv preprint arXiv:1403.4640*.

Gillani, N., Yasseri, T., Eynon, R., & Hjorth, I. (2014). Structural limitations of learning in a crowd: communication vulnerability and information diffusion in MOOCs. *Scientific Reports*, *4*, 6447.

Gorham, J. (1988). The relationship between verbal teacher immediacy behaviors and student learning. *Communication Education, 37,* 40-53

Gorham, J. (1988). The relationship between verbal teacher immediacy behaviors and student learning. *Communication Education, 37*, 40-53.

Green, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis 11*(3): 255–274.

Gunawardena, C. N. (1995). Social presence theory and implications for interaction and collaborative learning in computer conferences. *International Journal of Education Telecommunications, 1*(2), 147-166.

Gunawardena, C. N., Nolla, A. C., Wilson, P. L., Lopez-Islas, J. R., Ramirez-Angel, N., & Megchun-Alpizar, R. M. (2001). A cross-cultural study of group process and development in online conferences. *Distance Education, 22*(1), 85-121.

Gunawardena, C. N., & Zittle, F. J. (1997). Social presence as a predictor of satisfaction within a computer-mediated conferencing environment. *American Journal of Distance Education*, *11*(3), 8-26.

Guo, S., & Wu, W. (2015). Modeling student learning outcomes in MOOCs. In *The 4th International Conference on Teaching, Assessment, and Learning for Engineering* (pp. 1305-1313).

Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014). Attrition in MOOC: Lessons learned from drop-out students. *Learning Technology for Education in Cloud. MOOC and Big Data Communications in Computer and Information Science, 446*(2014), 37–48. http://dx.doi.org/10.1007/978-3-319-10671-7_4.

Hackman, M. Z., & Walker, K. B. (1990). Instructional communication in the televised classroom: The effects of system design and teacher immediacy on student learning and satisfaction. *Communication Education*, *39*(3), 196-206.

Hair, J.F. (1998), *Multivariate Data Analysis*, Prentice-Hall, Upper Saddle River, NJ.

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods.* University of California, Riverside. *CA*. http://faculty.ucr.edu/~hanneman/nettext/

Hecking, T., Chounta, I. A., & Hoppe, H. U. (2016, April). Investigating social and semantic user roles in MOOC discussion forums. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 198-207). ACM.

Hecking, T., Hoppe, H. U., & Harrer, A. (2015, August). Uncovering the structure of knowledge exchange in a MOOC discussion forum. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 1614-1615). IEEE.

Honeychurch, S., Bozkurt, A., Singh, L., & Koutropoulos, A. (2017). Learners on the periphery: Lurkers as invisible learners. *European Journal of Open, Distance and E-learning*, *20*(1), 192-212.

Hostetter, C., & Busch, M. (2006). Measuring up online: The relationship between social presence and student learning satisfaction. *Journal of Scholarship of Teaching and Learning, 6*(2), 1-12.

Houston II, S. L., Brady, K., Narasimham, G., & Fisher, D. (2017, April). Pass the idea please: The relationship between network position, direct engagement, and course performance in MOOCs. *Proceedings of the Fourth (2017) ACM Conference on Learning@Scale* (pp. 295-298). ACM.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hu J., Dowell N., Brooks C., Yan W. (2018) Temporal Changes in Affiliation and Emotion in MOOC Discussion Forum Discourse. In: Penstein Rosé C. et al. (Eds.), *Artificial Intelligence in Education* (pp. 145-149). Springer, Cham. Retrieve from: https://link.springer.com/chapter/10.1007/978-3-319-93846-2_26#citeas

Hughes, M., Ventura, S., & Dando, M. (2007). Assessing social presence in online discussion groups: A replication study. *Innovations in Education and teaching International, 44*(1), 17-29.

Jiang, S., Fitzhugh, S. M., & Warschauer, M. (2014). Social positioning and performance in MOOCs. *Proceedings of Graph-Based Educational Data Mining Workshop at the 7th International Conference on Educational Data Mining* (pp. 55-58). CEUR-WS.

Joksimović, S., Gašević, D., Kovanović, V., Riecke, B. E., & Hatala, M. (2015). Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning*, *31*(6), 638-654.

Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V., & De Kereki, I. F. (2016, April). Translating network position into performance: importance of centrality in different network configurations. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 314-323). ACM.

Jones, C. R., Ferreday, D., & Hodgson, V. (2008). Networked learning a relational approach: Weak and strong ties. *Journal of Computer Assisted Learning, 24*(2), 90–102. doi:10.1111/j.1365-2729.2007.00271.x

Jung, Y., & Lee, J. (2018). Learning engagement and persistence in Massive Open Online Courses (MOOCS). *Computers & Education, 122,* 9–22. https://doi.org/10.1016/j.compedu.2018.02.013.

Kamradt, T. F., & Kamradt, E. J. (1999). Structured design for attitudinal instruction. In C. M. Reigeluth (Eds.), *Instructional design theories and models: A new paradigm of instructional theory* (Vol. 2, pp. 563–590). Mahwah, NJ: Lawrence Erlbaum Associates.

Kang, M., & Im, T. (2013). Factors of learner–instructor interaction which predict perceived learning outcomes in online learning environment. *Journal of Computer Assisted Learning, 29*(3), 292– 301. doi:10.1111/jcal.12005

Kang, M., Liew, B.T., Kim, J. & Park, Y. (2014). Learning presence as a predictor of achievement and satisfaction in online learning environments. *International Journal on E-Learning, 13*(2), 193-208.

Ke, F. (2010). Examining online teaching, cognitive, and social presence for adult students. *Computers & Education*, *55*(2), 808-820.

Kearney, P., Plax, T. G., & Wendt-Wasco, N. J. (1985). Teacher immediacy for affective learning in divergent college classes. *Communication Quarterly*, *33*(1), 61-74.

Kellogg, S., Booth, S., & Oliver, K. (2014). A social network perspective on peer supported learning in MOOCs for educators. *The International Review of Research in Open and Distributed Learning*, *15*(5).

Kelly, D., & Gorham, J. (1988). Effects of immediacy on recall of information. *Communication Education, 37*, 198-207.

Kennedy, G., Coffrin, C., De Barba, P., & Corrin, L. (2015, March). Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance. *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 136-140).

Khalil, M., & Ebner, M. (2016). What massive open online course (MOOC) stakeholders can learn from learning analytics?. *ArXiv preprint arXiv:1606.02911*. https://doi.org/10.1007/978-3-319-17727-4_3-1

Kilgore, W., & Lowenthal, P. R. (2015). The human element MOOC: An experiment in social presence. In R. D. Wright (Ed.), *Student-teacher interaction in online learning environments* (pp. 373–391). Hershey, PA: IGI Global.

Kim, W., Watson, S. L., & Watson, W. R. (2016). Perceived learning in three MOOCs targeting attitudinal change. *Educational Media International*, *53*(3), 168-183.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170-179). ACM.

Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, *355*(6322), 251-252.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014, October). Predicting MOOC dropout over weeks using machine learning methods. *Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs* (pp. 60-65). Retrieved from: http://www2.informatik.huberlin.de/~kloftmar/publications/emnlp_mooc.pdf

Knox, J. (2014). Digital culture clash:"massive" education in the E-learning and Digital Cultures MOOC. *Distance Education*, *35*(2), 164-177.

Kohlmeyer, J. M., Seese, L. P., & Sincich, T. (2011). Online versus traditional accounting degrees: Perceptions of public accounting professionals. In *Advances in accounting education: Teaching and curriculum innovations* (pp. 139-165). Emerald Group Publishing Limited.

Kollock, P., & Smith, M. (1996). Managing the virtual commons. *Computer-Mediated Communication: Linguistic, Social, and Cross-cultural Perspectives*, 109-128.

Kop, R. (2011). The challenges to connectivist learning on open online networks: Learning experiences during a massive open online course. *International Review of Research in Open and Distributed Learning*, *12*(3), 19-38.

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016, April). Towards automated content analysis of discussion transcripts: A cognitive presence case. *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 15-24). ACM.

Kozan, K., & Richardson, J. C. (2014). Interrelationships between and among social, teaching, and cognitive presence. *The Internet and higher education*, *21*, 68-73.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Leh, A. S. (2001). Computer-mediated communication and social presence in a distance learning environment. *International Journal of Educational Telecommunications, 7*(2), 109–128.

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015, July). Support vector machines and word2vec for text classification with semantic features. *Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing* (ICCI* CC) (pp. 136-140). IEEE.

Liu, W., Kidziński, Ł., & Dillenbourg, P. (2016). Semi-automatic annotation of MOOC forum posts. In *State-of-the-Art and Future Directions of Smart Learning* (pp. 399-408). Springer, Singapore.

Liu, X., Liu, S., Lee, S., & Magjuka, R. J. (2010). Cultural differences in online learning: International student perceptions. *Educational Technology & Society*, *13*(3), 177-188.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Sun, C., Lin, L., & Wang, X. (2016). Learning natural language inference using bidirectional LSTM model and inner-attention. *arXiv preprint arXiv:1605.09090*.

Liu, M., Zou, W., Li, C., Shi, Y., Pan, Z., & Pan, X. (2019). Using learning analytics to examine relationships between learners' usage data with their profiles and perceptions: a case study of a MOOC designed for working professionals. In *Utilizing Learning Analytics to Support Study Success* (pp. 275-294). Springer, Cham.

Liu, M., Zou, W., Shi, Y., Pan, Z., & Li, C. (2019). What do participants think of today's MOOCs: an updated look at the benefits and challenges of MOOCs designed for working professionals. *Journal of Computing in Higher Education*, 1-23.

Lowenthal, P. R. (2009). Social presence. In *Encyclopedia of Distance Learning, Second Edition* (pp. 1900-1906). IGI Global.

Lowenthal, P. R. (2010). Social presence. In *Social Computing: Concepts, Methodologies, Tools, and Applications* (pp. 129-136). IGI Global.

Lu, D., & Farzan, R. (2015, June). Time to Introduce myself! impact of self-disclosure timing of Newcomers in Online Discussion Forums. *Proceedings of the ACM Web Science Conference* (pp. 1-9). Retrieved from: https://www.researchgate.net/profile/Rosta_Farzan/publication/278965110_Time_to_Introduce_Myself_Impact_of_Self-disclosure_Timing_of_Newcomers_in_Online_Discussion_Forums/links/56e96fb208ae47bc651c736c/Time-to-Introduce-Myself-Impact-of-Self-disclosure-Timing-of-Newcomers-in-Online-Discussion-Forums.pdf

Lu, J., & Churchill, D. (2014). The effect of social interaction on learning engagement in a social networking environment. *Interactive learning environments*, *22*(4), 401-417.

Ludwig-Hardman, S., & Dunlap, J. C. (2003). Learner support services for online students: Scaffolding for success. *The International Review of Research in Open and Distributed Learning*, *4*(1).

McKlin, T., Harmon, S. W., Evans, W., & Jones, M. J. (2001, November). *Cognitive Presence in Web-Based Learning: A Content Analysis of Student's Online Discussions.* Paper Presented at the National Convention of the Association for Educational Communications and Technology, Atlanta, GA.

McLellan, H. (1999). Online education as interactive experience: Some guiding models. *Educational technology*, *39*(5), 36-42.

McLoughlin, C., & Oliver, R. (2000). Designing learning environments for cultural inclusivity: A case study of indigenous online learning at tertiary level. *Australasian Journal of Educational Technology*, *16*(1).

Mehrabian, A. (1969). Significance of posture and position in the communication of attitude and status relationships. *Psychological Bulletin*, *71*(5), 359.

Mehrabian, A. (1981), *Silent messages: Implicit communication of emotions and attitudes* (2nd ed.). Belmont, CA Wadsworth.

Meyer, K. A. (2004). Putting the distance learning comparison study in perspective: Its role aspersonal journey research. *Online Journal of Distance Learning Administration, 7*(1). Retrieved from http://www.westga.edu/~distance/ojdla/spring71/meyer71.pdf

Moon, S., Potdar, S., & Martin, L. (2014). Identifying student leaders from MOOC discussion forums through language influence. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 15–20). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/v1/w14-4103

Moore, R. L., Oliver, K. M., & Wang, C. (2019). Setting the pace: examining cognitive processing in MOOC discussion forums with automatic text analysis. *Interactive Learning Environments*, 1-15.

Morris, M., & Ogan, C. (1996). The Internet as mass medium. *Journal of Computer-Mediated Communication*, *1*(4), JCMC141.

Morris, N. P., Swinnerton, B. J., & Hotchkiss, S. (2015). Can demographic information predict MOOC learner outcomes?. *Proceedings of the European MOOC Stakeholder*. Leeds.

Neelen, M., & Fetter, S. (2010). Lurking: A challenge or a fruitful strategy? A comparison between lurkers and active participants in an online corporate community of practice. *International Journal of Knowledge and Learning, 6*(4), 269–284.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577-8582.

Nistor, N., Dascalu, M., Serafin, Y., & Trausan-Matu, S. (2018). Automated dialog analysis to predict blogger community response to newcomer inquiries. *Computers in Human Behavior*, *89*, 349-354.

Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Nonnecke, R. B. (2000). Lurking in email-based discussion lists. [Doctoral dissertation, South Bank University]. Retrieve from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.5694&rep=rep1&type=pdf

Nonnecke, B., Andrews, D., & Preece, J. (2006). Non-public and public online community participation: Needs, attitudes and behavior. *Electronic Commerce Research, 6*(1), 7–20.

Nonnecke, B., & Preece, J. (2001). Why lurkers lurk. *Proceedings of Americas Conference on Information Systems* (pp. 1–10). Retrieved from: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1733&context=amcis2001

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315-341.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543). Retrieve from: https://www.aclweb.org/anthology/D14-1162.pdf

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Piaget, J. (2013). *The construction of reality in the child* (Vol. 82). Routledge.

Picciano, A. (2002). Beyond student perceptions: issues of interaction, presence, and performance in an online course. *Journal of Asynchronous Learner Networks, 6*(1), 21- 40.

Poquet, P., & Dawson, D. (2016, April). Untangling MOOC learner networks. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 208-212). ACM.

Qu, H., & Chen, Q. (2015). Visual analytics for MOOC data. *IEEE Computer Graphics and Applications, 35*(6), 69–75.

Ramesh, A., Goldwasser, D., Huang, B., Daume, H., & Getoor, L. (2014, June). Understanding MOOC discussion forums using seeded LDA. *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 28-33). ACM.

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier.* MIT Press.

Richardson, J., & Swan, K. (2003). Examining social presence in online courses in relation to students' perceived learning and satisfaction. *Online Learning, 7*(1). http://dx.doi.org/10.24059/olj.v7i1.1864

Rohs, M., & Ganz, M. (2015). MOOCs and the claim of education for all: A disillusion by empirical data. *International Review of Research in Open and Distributed Learning*, *16*(6), 1-19.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458-472.

Rong, X. (2014). Word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rossi, L. A., & Gnawali, O. (2014, August). Language independent analysis and classification of discussion threads in Coursera MOOC forums. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)* (pp. 654-661). IEEE.

Rourke, L., & Anderson, T. (2002). Exploring social communication in computer conferencing. *Journal of Interactive Learning Research*, *13*(3), 259-275.

Rourke, L., Anderson, T., Archer, W., & Garrison, D. R. (1999). Assessing social presence in asynchronous text-based, computer conferencing. *Journal of Distance Education, 14*(3), 51–70.

Russo, T. C., & Koesten, J. (2005). Prestige, centrality, and learning: A social network analysis of an online class. *Communication Education*, *54*(3), 254-261.

Rovai, A. P. (2002). Sense of community, perceived cognitive learning, and persistence in asynchronous learner networks. *The Internet and Higher Education*, *5*(4), 319-332.

Ryan, A., & Tilbury, D. (2013). Flexible Pedagogies: new pedagogical ideas. *Higher Education Academy,* London. Retrieved from: https://www.heacademy.ac.uk/system/files/resources/npi_report.pdf

Sanders, J., & Wiseman, R. (1990). The effects of verbal and nonverbal teacher immediacy on perceived cognitive, affective, and behavioral learning in the multicultural classroom. *Communication Education, 39*, 341-353.

Scott, J. (1991). *Social network analysis: A handbook*. London: Sage.

Scott, M. D., & Wheeless, L. R. (1977). Communication apprehension, student attitudes, and levels of satisfaction. *Western Journal of Communication, 41*, 188–198.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., ... & Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media*, *1*(2).

Shea, P., Hayes, S., Vickers, J., Gozza-Cohen, M., Uzuner, S., Mehta, R., ... & Rangan, P. (2010). A re-examination of the community of inquiry framework: Social network and content analysis. *The Internet and Higher Education*, *13*(1-2), 10-21.

Shi, L., Cristea, A. I., Toda, A. M., & Oliveira, W. (2020, June). Exploring Navigation Styles in a FutureLearn MOOC. *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 45-55). Springer, Cham.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. John Wiley & Sons.

Simonson, M. R. (1979). Designing instruction for attitudinal outcomes. *Journal of Instructional Development, 2*, 15–19.

Simonson, M. R., & Maushak, N. (1996). Situated learning, instructional technology, and attitude change. In H. McLellan (Eds.), *Situated Learning Perspectives* (pp. 225–242). Englewood Cliffs: Educational Technology Publications Inc.

Sosniak, L. A. (1994). *Bloom's taxonomy*. L. W. Anderson (Ed.). Chicago, IL: Univ. Chicago Press.

Stiller, K. D., & Bachmaier, R. (2017). Dropout in an online training for trainee teachers. *European Journal of Open, Distance and e-Learning*, *20*(1), 80-95.

Strong, R., Irby, T. L., Wynn, J. T., & McClure, M. M. (2012). Investigating Students' Satisfaction with eLearning Courses: The Effect of Learning Environment and Social Presence. *Journal of Agricultural Education, 53*(3).

Stump, G. S., DeBoer, J., Whittinghill, J., & Breslow, L. (2013). Development of a framework to classify MOOC discussion forum posts: Methodology and challenges. In *NIPS Workshop on Data Driven Education* (pp. 1-20).

Swan, K., & Shih, L. F. (2005). On the nature and development of social presence in online course discussions. *Journal of Asynchronous Learner Networks*, *9*(3), 115-136.

Szeto, E. (2015). Community of Inquiry as an instructional approach: What effects of teaching, social and cognitive presences are there in blended synchronous learning and teaching? *Computers & Education*, *81*, 191-201.

Tabachnick, B. G., & Fidell, L. S. (2001). Principal components and factor analysis. *Using Multivariate Statistics, 4*(1), 582-633.

Tashakkori, A., & Teddlie, C. (2003). *Handbook on mixed methods in the behavioral and social sciences*. Sage Publications, Thousand Oaks, CA.

Tashakkori, A., Teddlie, C., & Teddlie, C. B. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). Sage Publications, Thousand Oaks, CA.

Tedjamulia, S.J., Dean, D.L., Olsen, D.R., & Albrecht, C.C. (2005). Motivating content contributions to online communities: Toward a more comprehensive theory. *Proceedings of the 38th Hawaii International Conference on System Sciences* (HICSS). IEEE.

Tseng, S. F., Tsao, Y. W., Yu, L. C., Chan, C. L., & Lai, K. R. (2016). Who will pass? Analyzing learner behaviors in MOOCs. *Research and Practice in Technology Enhanced Learning, 11*(1), 8. https://doi.org/10.1186/s41039-016-0033-5

Tu, C. H. (2000). On-line learning migration: from social learning theory to social presence theory in a CMC environment. *Journal of Network and Computer Application, 23*(1), 27–37.

Van der Hulst, R. C. (2009). Introduction to Social Network Analysis (SNA) as an investigative tool. *Trends in Organized Crime*, *12*(2), 101-121.

Van Mierlo, T. (2014). The 1% rule in four digital health social networks: An observational study. *Journal of Medical Internet Research, 16*(2).

Vaquero, L. M., & Cebrian, M. (2013). The rich club phenomenon in the classroom. *Scientific Reports*, *3*, 1174.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wasko, M. M., Teigland, R., & Faraj, S. (2009). The provision of online public goods: Examining social structure in an electronic network of practice. *Decision support systems*, *47*(3), 254-265.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Wen, M., Yang, D., & Rosé, C. P. (2014). Sentiment analysis in MOOC discussion forums: What does it tell us. *Proceedings of the 7th international conference on educational data mining* (EDM 2014) (pp. 130–137). Retrieved from: http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/130_EDM-2014-Full.pdf

Whiteman, J. A. M. (2002). *Interpersonal communication in computer mediated learning*. ERIC Document Reproduction Service, No. ED 465 977. Retrieved from: http://eric.ed.gov/?id=ED465997

Natural language processing. (2020, August). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Natural_language_processing

Wise, A. F., & Cui, Y. (2018). Learning communities in the crowd: Characteristics of content related interactions and social relationships in MOOC discussion forums. *Computers & Education*, *122*, 221-242.

Wise, A. F., Cui, Y., & Jin, W. Q. (2017, March). Honing in on social learner networks in MOOC forums: Examining critical network definition decisions. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 383-392). ACM.

Wise, A. F., Cui, Y., & Vytasek, J. (2016, April). Bringing order to chaos in MOOC discussion forums with content-related thread identification. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 188-197). ACM.

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. *International Educational Data Mining Society*.

Wong, J. S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015, March). An analysis of MOOC discussion forum interactions from the most active users. *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 452-457). Springer, Cham.

Wong, A. W., Wong, K., & Hindle, A. (2019). Tracing Forum Posts to MOOC Content using Topic Analysis. *arXiv preprint arXiv:1904.07307*.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal prediction of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, *58*, 119-129.

Xing, W., Tang, H., & Pei, B. (2019). Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of MOOCs. *The Internet and Higher Education,* 100690.

Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An Improved Random Forest Classifier for Text Categorization. *Journal of Computers, 7*(12), 2913-2920.

Yang, D., Wen, M., Howley, I., Kraut, R., & Rose, C. (2015, March). Exploring the effect of confusion in discussion forums of massive open online courses. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 121-130). ACM.

Yusof, N., & Rahman, A. A. (2009, April). Students' interactions in online asynchronous discussion forum: A Social Network Analysis. *Proceedings of 2009 International Conference on Education Technology and Computer* (pp. 25-29). IEEE.

Feilzer, M.Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of mixed methods research*, *4*(1), 6-16.

Zhang, Q., Peck, K. L., Hristova, A., Jablokow, K. W., Hoffman, V., Park, E., & Bayeck, R. Y. (2016). Exploring the communication preferences of MOOC learners and the value of preference-based groups: Is grouping enough?. *Educational Technology Research and Development*, *64*(4), 809-837.

Zhang, W., & Storck, J. (2001). Peripheral members in online communities. *Proceedings of the Americas Conference on Information Systems* (p.7). Retrieved from: http://aisel.aisnet.org/amci2001/117

Zhang, M., Yin, S., Luo, M., & Yan, W. (2017). Learner control, user characteristics, platform difference, and their role in adoption intention for MOOC learning in China. *Australasian Journal of Educational Technology*, *33*(1).